UNIVERSIDAD CARLOS III DE MADRID

TESIS DOCTORAL

Estimating Non-linear Models with Applications to Health, Labor and Education Economics

Autora:
Alejandra Traferri

Directora:
Raquel Carrasco Perea

DEPARTAMENTO DE ECONOMÍA

Getafe, Junio de 2011

TESIS DOCTORAL

Estimating Non-linear Models with Applications to Health, Labor and Education Economics

Autora: Alejandra Traferri

Directora: Raquel Carrasco Perea

Firma del Tribunal Calificador:

Firma

Presidente:

Vocal:

Vocal:

Vocal:

Secretario:

Calificación:

Getafe, de de

A mis abuelos.

Contents

Acknowledgements

Quisiera agradecer especialmente a mi directora, Raquel Carrasco, por todo lo que me ha enseñado a lo largo de este proceso, por saber encarrilarme cuando me he desviado del objetivo y por alentarme cuando lo he hecho bien. Le agradezco la confianza y la paciencia que ha tenido conmigo.

Gracias al Departamento de Economía por abrirme las puertas de esta universidad, por la formación que me han dado y por brindarme todo lo necesario para realizar mi trabajo lo mejor posible. En particular, quisiera agradecer a mi coautor, Jesús Carro, del cual he aprendido muchísimo y a Juanjo Dolado que desde el primer día ha estado dispuesto a ayudarme. Les agradezco la generosidad que han tenido al brindarme su tiempo y al compartir sus ideas conmigo. También quiero agradecer especialmente a Paco Marmol, una persona luchadora y un profesor ejemplar. Me siento afortunada de haber podido disfrutar de sus clases.

Gracias al Institute for Economic Development (IED) de Boston University y a todos los profesores, por recibirme y tratarme como a una estudiante más, en especial a Dilip Mookherjee. Y por supuesto, un cariñoso agradecimiento a mis compañeras de oficina y charlas, Marian y Fernanda.

Cómo no agradecer a mis compañeros del doctorado, con quiénes hemos compartido, disfrutado y sufrido cada momento de esta etapa. De cada uno de ellos he aprendido algo. En especial, quisiera agradecer al Córdoba team formado por Nieves, Rodo y Gastón, ya que desde el primer momento que desembarcamos en Barajas hasta el final hemos estado luchando y sobreviviendo juntos, y a Jess, Jou, y Abelardo, por brindarme su amistad y por los momentos únicos que hemos disfrutado.

A la comunidad argentina en Getafe, que crecía año a año (razón por la que no los nombraré uno a uno, pero va para cada uno de ellos sin excepción), les doy las gracias por hacer que estos años fuesen mucho más llevaderos y que por momento

sintiera que estaba en casa. Se me vienen a la memoria tantos recuerdos, tantos momentos lindos vividos, que lo único que puedo decir es que siempre estarán en mi corazón y que me llevo algo de cada uno de ellos. Espero que siempre haya un Getafe donde reencontrarnos y seguir compartiendo momentos.

A mi mamá, mi papá y mis 2 hermanas, porque sé que han vivido cada año del doctorado como si lo estuviesen haciendo ellos mismos, y porque cada vez que los visitaba me llenaban de abrazos y amor, y me preparaban todos esos sabores que uno siempre extraña cuando está lejos, y que en ningún lado saben como en casa. Gracias a ellos siempre volvía con fuerza y energía para continuar.

A Gastón, porque su apoyo y su amor sin límites han sido el aliento necesario para terminar esto entera. Porque ha creído en mí incluso cuando yo dejaba de hacerlo. Porque siempre trató de contagiarme su pasión por lo que hacemos para mantenerme ilusionada. Porque ha sido parte de mi cerebro. Y simplemente porque él ha sido y es todo para mí.

A todos, a los que he nombrado y a los que me he podido olvidar, gracias... hasta el infinito y más allá.

Introduction

This dissertation is composed of three studies of non-linear econometric models, with applications to Health, Labor and Education Economics.

Chapter 1 studies the differences in the proportion of temporary employees of domestic and foreign firms in the Spanish manufacturing sector. The objective of the chapter is to determine if, after controlling for a large set of observable firm characteristics and unobservable firm-specific time-invariant components, there is still a relationship between firm nationality and the type of employment contracts that firms offer.

For that purpose, I estimate standard censored Tobit and Heckman sample selection models (also known as type I Tobit and Heckman two-step models, respectively), using data from the Survey of Managerial Strategies (Encuesta sobre Estrategias Empresariales, ESEE), which includes a representative number of Spanish firms in the manufacturing sector during 1991-2005.

The estimations show that firm nationality has a significant effect over the probability that firms hire temporary workers and over the proportion of temporary workers for firms that choose to hire temporary workers. The size and significance of the effects depend on firm size. In the case of the Heckman estimations, for example, a higher proportion of foreign capital implies a lower average probability of hiring temporary workers for small and medium firms, but not for large firms. Likewise, a higher proportion of foreign capital implies a decrease in the average proportion of temporary workers (for firms who choose to hire temporary workers) for medium and large firms, but not for small firms.

Chapter 1 provides two contributions to the literature. First, it presents a further account of the differences between domestic and foreign firms. In particular, I show that domestic and foreign firms do not differ only in wages and productivity, but also on the types of labor contracts that they offer. Second, the chapter provides a detailed study of the determinants of temporary employment in the

Spanish manufacturing sector, focusing in particular on firm nationality and firm size. My findings indicate that a labor reform trying to reduce the use of fixed-term employment contracts should provide different incentives to firms of different size.

Chapter 2 considers the estimation of a dynamic ordered probit with fixed effects, and its application to the study of the determinants of self-assessed health status (SAH).[1]

SAH has been used as a proxy for true overall individual health status in many socioeconomic studies. Moreover, it has been shown to be a good predictor of mortality and of subsequent demand of medical care (see for example van Doorslaer, Koolman and Jones 2004).

Contoyannis, Jones and Rice (2004) studied the dynamics and effects of socioeconomic variables on SAH for the British Household Panel Survey, by performing a random effects analysis on a dynamic ordered probit model. Among other aims, they tried to determine the relative contribution of state dependence and unobserved heterogeneity in explaining the observed persistence in SAH.

This chapter applies a 'fixed effects' approach instead, which allows us to leave unrestricted the joint distribution of the two individual effects and their correlation with the explanatory variables, and to avoid the initial conditions problem.

In addition to accounting for unobserved factors that affect health status (index shifts), it is also important to take into account the possible heterogeneity in reporting behavior (cut-point shifts), which may occur if individuals use different thresholds when assessing their health and reporting it in the SAH categorical variable (i.e. two individuals may report a different value of SAH even though they have the same level of true health).

Despite the advantages of fixed effects over random effects estimations, there have been only few applications in non-linear models in health economics (as can be seen by reading Jones's 2007 handbook chapter). This is due to the difficulty in dealing with the incidental parameters problem, which is specially severe in the model studied in the chapter because it contains two fixed effects (index and cut-point shifts).

To account for the incidental parameters problem, we apply Carro's (2007) approach to bias reduction. We compare the resulting estimators with the 'standard'

[1]This chapter is based on Carro and Traferri (2009).

Maximum Likelihood estimators, and with the bias-corrected estimators proposed by Bester and Hansen (2009). In Montecarlo simulations, we show that both Maximum Likelihood and Bester and Hansen's estimators under-perform relative to the estimators obtained by following Carro's approach. Moreover, we also find that large biases remain in the case of Bester and Hansen's estimator, even for relatively large panel sizes.

We estimate the model using the British Household Survey in the period 1991-2006. Based on our best estimates, the two fixed effects exhibit important variation so it is relevant to account for both when estimating the effect of other variables. Our estimates also show that state dependence is very important even though we have controlled for unobserved heterogeneity and some forms of objective health measures. The latter are the variables with higher marginal effects.

The contributions of this chapter are twofold. First, the chapter contributes to the recent literature on bias correction in nonlinear panel data models by applying and studying the finite sample properties of two of the existing proposals to the ordered probit case. We find that the most direct and easily applicable correction to our model is not the best one and still has important biases in our sample sizes. Second, we contribute to the literature that studies the determinants of Self-Assessed Health measures by applying the previous analysis on estimation methods to the British Household Panel Survey.

Finally, in Chapter 3, I study gender differences in major choice and college entrance probabilities in the University of Campinas, a Brazilian public university dependent on the State of São Paulo.

As with most Brazilian public universities, candidates for entry into University of Campinas first select a major, and then compete for a place in that major by taking a major-specific entrance exam. This singular characteristic of the Brazilian case allows me to differentiate the effect of gender on major-specific entrance probabilities and preferences.

I propose a model and econometric strategy which can account for two important issues, selectivity bias and the fact that expected utility depends on the probability of entering the different majors. I estimate this model using a novel dataset obtained from University of Campina's Permanent Commission for Vestibular Exams (Comvest).

I find evidence of gender differences in preferences and entrance probabilities. For most majors, gender differences in major choice are mostly explained by differences in preferences. However, for the most demanding majors (those that require higher grades from students), differences in major choice are explained in a large proportion by differences in entrance probabilities. Finally, I find that gender has important interactions with other variables. In particular, gender effects depend on education, socioeconomic characteristics and family background.

This chapter has three contributions. First, the econometric model is able to account for selection bias, in contrast to previous papers, which have assumed that the errors of the probability of entry and choice equations were independent, i.e. there was no selectivity bias by assumption (see for example, Montmarquette et al. 2002). Second, the chapter introduces a novel database, which can be used to disentangle the differential effects of probability of entry and preferences in gender differences in major choice. Third, this chapter provides the first detailed analysis of the determinants of major choice in Brazil. A few papers analyze the determinants of performance in Entrance Test Exams (Guimaraes and Sampaio 2007, 2008, Calvacanti et al. 2009), but the choice of major has not been analyzed in detail. Performing such analysis is important because of its possible relation with gender inequality, which is an important issue for the Brazilian Federal Government. For example, the Federal Government has recently introduced over 400 projects directed at enhancing equal opportunities for men and women, which will be performed by 22 government institutions between 2008 and 2011 (Pinheiro et al. 2008).

Resumen

Esta tesis se compone de tres trabajos de investigación en las áreas de la Economía de la Salud, Trabajo y Educación.

El capítulo 1 estudia las diferencias en la proporción de empleados temporales entre las empresas nacionales y extranjeras del sector manufacturero español. El objetivo del capítulo es determinar si existe una relación entre la nacionalidad de las empresas y el tipo de contratos de trabajo ofrecidos, aún después de controlar por un amplio conjunto de características observables y componentes no observables que no varían en el tiempo.

A tal efecto, estimo modelos Tobit censurados, y modelos de selección de Heckman (también conocidos como Tobit de tipo I y modelo Heckman de dos etapas, respectivamente), utilizando datos de la Encuesta sobre Estrategias Empresariales (ESEE), que incluye un número representativo de empresas españolas en el sector manufacturero durante 1991-2005.

Las estimaciones muestran que la nacionalidad de la empresa tiene un efecto significativo sobre la probabilidad de que las empresas contraten trabajadores temporales y sobre la proporción de trabajadores temporales, para las empresas que optan por contratar a trabajadores temporales. El tamaño y significancia de los efectos dependen del tamaño de la empresa. En el caso del modelo Heckman, por ejemplo, una mayor proporción de capital extranjero implica una menor probabilidad media de contratación de trabajadores temporales para pequeñas y medianas empresas, pero no para empresas grandes. Del mismo modo, una mayor proporción de capital extranjero implica una disminución de la proporción media de trabajadores temporales (para las empresas que optan por contratar a trabajadores temporales) para las empresas medianas y grandes, pero no para las pequeñas empresas.

El capítulo 1 tiene dos contribuciones principales. En primer lugar, el capítulo presenta un análisis adicional de las diferencias entre empresas nacionales y extranjeras. En particular, se muestra que las empresas nacionales y extranjeras no sólo se diferencian en productividad y nivel de salarios, sino también en los tipos de contrato laborales que ofrecen a sus trabajadores. En segundo lugar, el capítulo ofrece un estudio detallado de los determinantes del empleo temporal en el sector manufacturero español, prestando especial atención a la nacionalidad y tamaño de las empresas. Los resultados indican que una reforma laboral que trate de reducir el uso de contratos laborales temporales deben ofrecer incentivos distintos a empresas de diferente tamaño.

El capítulo 2 estudia la estimación de un modelo probit ordenado dinámico con efectos fijos, y su aplicación al estudio de los determinantes del estado de salud autoreportado (ESA).[2]

El ESA se ha utilizado como sustituto del verdadero estado de salud en numerosos estudios socioeconómicos. Por otra parte, también se ha demostrado que es un buen predictor de la mortalidad y de la demanda de atención médica (véase, por ejemplo van Doorslaer, Koolman y Jones 2004).

Contoyannis, Jones y Rice (2004) estudiaron la dinámica y los efectos de variables socioeconómicas sobre la ESA para la Encuesta de Panel de Hogares Británica (British Household Panel Survey). Concretamente, los autores estudiaron un modelo probit ordenado dinámico con efectos aleatorios (random effects). Entre otros objetivos, Contoyannis, Jones y Rice buscaban determinar la contribución relativa de la dependencia del estado (state dependence) y la heterogeneidad no observada en la persistencia observada en la ESA.

En este capítulo, aplicamos un análisis de efectos fijos, en vez de efectos aleatorios, lo que nos permite dejar libre la distribución conjunta de los dos efectos individuales y su correlación con las variables explicativas, así como evitar el problema de las condiciones iniciales.

Además de tener en cuenta a los factores no observados que afectan el estado de salud (cambios de índice, index shifts), también tomamos en cuenta la posible heterogeneidad en los criterios de reporte del estado de salud (cambios de punto de corte, cut point shifts), que puede producirse si los individuos utilizan distintos

[2]Este capítulo está basado en Carro y Traferri (2009).

umbrales para evaluar su salud (es decir, dos personas pueden informar de un valor diferente para la ESA aunque tengan el mismo nivel de salud verdadera).

A pesar de las ventajas de efectos fijos sobre los efectos aleatorios, ha habido pocas aplicaciones en modelos no lineales en el área de la economía de la salud (tal como puede verse al leer el capítulo del Handbook of econometrics de Jones 2007). Esto se debe a la dificultad de lidiar con el problema de los parámetros incidentales, el cual es especialmente grave en nuestro modelo, ya que contiene dos efectos fijos (index y cut point shifts).

Para tener en cuenta el problema de los parámetros incidentales, aplicamos el enfoque de reducción de sesgo de Carro (2007), y comparamos los estimadores resultantes con los estimadores "estándar" de máxima verosimilitud, y con los estimadores corregidos de Bester y Hansen (2009). Las simulaciones de Montecarlo muestran que tanto los estimadores de máxima verosimilitud como los de Bester y Hansen tienen un mayor sesgo que los obtenidos bajo el enfoque de Carro. Por otra parte, también encontramos que los estimadores de Bester y Hansen tienen un gran sesgo, incluso para paneles relativamente grandes.

Estimamos el modelo propuesto usando datos de la Encuesta de Panel de Hogares Británica en el período 1991-2006. Nuestras mejores estimaciones muestran que los dos efectos fijos presentan una variación significativa para distintos individuos, por lo que es importante tener en cuenta ambos efectos al estimar el efecto de otras variables. Nuestras estimaciones muestran también que la dependencia de estado es muy importante, a pesar de que se haya controlado por la heterogeneidad no observada y algunas medidas objetivas de salud. Estas últimas variables son las que poseen los mayores efectos marginales.

Este capítulo tiene dos contribuciones principales. En primer lugar, el capítulo contribuye a la literatura reciente sobre la corrección de sesgo en los datos de panel de modelos no lineales, mediante la aplicación y el estudio de las propiedades de muestra finita de dos de las propuestas existentes para el caso probit ordenado. Encontramos que la corrección más directa y de fácil aplicación a nuestro modelo (la de Bester y Hansen) no es la mejor, y todavía tiene mantiene sesgos importantes para tamaños de panel como el que nosotros utilizamos. En segundo lugar, el capítulo contribuye a la literatura que estudia los determinantes de las medidas de salud autoreportada, mediante la aplicación del análisis anterior a la Encuesta de Panel de Hogares Británica.

Por último, en el capítulo 3 estudio las diferencias de género en la elección de carrera y en la probabilidad de ingreso a la universidad en la Universidad de Campinas, una universidad pública brasileña dependiente del Estado de São Paulo.

Como con la mayoría de las universidades públicas de Brasil, los candidatos para la entrada en la Universidad de Campinas eligen primero la carrera a la que desean entrar, y luego compiten por un lugar en esa carrera tomando un examen específico para dicha carrera. Esta característica singular del caso de Brasil me permite diferenciar el efecto del género sobre las probabilidades de entrada en cada carrera, y sobre las preferencias.

En este capítulo, propongo un modelo econométrico que tiene en cuenta dos problemas importantes: el sesgo de selección, y el hecho de que la utilidad esperada de una carrera depende de la probabilidad de entrada. Luego, estimo el modelo utilizando una novedosa base de datos obtenida de la Comisión Permanente de los Vestibulares (Comvest) de la Universidad de Campinas.

Las estimaciones proveen evidencia de que existen diferencias de género tanto en las preferencias, como en las probabilidades de entrada a las distintas carreras. Para la mayoría de las carreras, la mayor parte de las diferencias de género en la elección de las carreras es explicada por diferencias en preferencias. Sin embargo, para las carreras más exigentes (aquellas que requieren mayores calificaciones para entrar), las diferencias en la elección de carrera también son explicadas en gran medida por las diferencias en las probabilidades de entrada. Por último, también encuentro que el género tiene interacciones importantes con otras variables. En particular, las diferencias de género dependen de las variables de educación, las características socioeconómicas y los antecedentes familiares.

Este capítulo tiene tres contribuciones principales. En primer lugar, el capítulo presenta un modelo econométrico que tiene en cuenta el sesgo de selección, a diferencia de trabajos anteriores, que han asumido que los errores de las ecuaciones que determinan la probabilidad de entrada y la elección de los estudiantes son independientes, por lo que suponen que no hay sesgo de selectividad (véase, por ejemplo, Montmarquette et al. 2002). En segundo lugar, el capítulo introduce una nueva base de datos, que permite separar los efectos diferenciales de la probabilidad de entrada y las preferencias en las diferencias de género en la elección de carrera universitaria. En tercer lugar, este capítulo proporciona el primer análisis

detallado de las diferencias de género en la elección de carreras en Brasil. Dicho análisis es importante por los posibles efectos de las diferencias en la elección de carrera sobre las posteriores diferencias profesionales entre hombres y mujeres. Esta línea de investigación es muy importante para el Gobierno Federal de Brasil, que ha introducido más de 400 proyectos dirigidos a mejorar la igualdad de oportunidades entre hombres y mujeres, los que serán llevados a cabo por 22 instituciones gubernamentales 2008 y 2011 (Pinheiro et al. 2008).

CHAPTER 1

Temporary Employment and Firm Ownership Nationality: Evidence from Spain

Abstract. This paper analyzes the differences in the proportion of temporary employees of domestic and foreign firms in the Spanish manufacturing sector. I estimate sample selection models using data from the Survey on Managerial Strategies (ESEE) in 1991-2005. I find there is a clear relation between the nationality of the owners of the firm and the type of labor contracts offered, even after controlling for observable firm characteristics and unobservable fixed effects. In particular, the share of temporary employees is significantly lower for foreign firms and this effect decreases with firm size.

1. Introduction

The high share of temporary workers in total employment in Spain since the mid-1980s raised concern among policy-makers because of the potential negative effects of temporary employment on efficiency and equality. As has been noted in the literature, the co-existence of permanent and temporary contracts creates a segmented labor market, which may imply lower investment in human capital, unequal distribution of unemployment duration, lower mobility and larger wage dispersion (see, for example, Bentolila and Dolado 1994, Güell 2000, Garibaldi and Mauro 2002, Blanchard and Landier 2002).[1]

A series of employment legislation reforms were performed in 1994, 1997 and 2001, with the objective of reducing the share of temporary workers. The reforms lowered the hiring and dismissal costs of permanent workers and restricted the use of fixed-term contracts. According to the Index of Strictness of Employment

[1]The high share of temporary workers was a consequence of the 1984 Employment Protection Legislation Reform, which liberalized fixed-term contracts, in order to reduce the high unemployment rate. As a result, the share of temporary workers increased from 15.6% in 1987 to 33.65% in 1994. However, the reform had little effect on the unemployment rate, which was on average 19.41% between 1984 and 1994.

Protection Legislation of the OECD (Table A.1 in Appendix A), the regulation of permanent contracts in Spain was one of the most stringent regulations in the OECD in 1990, and the regulation of temporary contracts was one of the weakest regulations. After the reforms, the regulation of permanent contracts became more flexible and the regulation of temporary contracts strengthened, becoming one of the most stringent by 2003.

Nevertheless, these reforms had no impact on the share of temporary workers, which kept above 30% from 1994 to 2007. In comparison with other OECD countries, Spain still has one of the largest shares of temporary employment, which is 20 percentage points above the OECD and European means (see Table A.2 in Appendix A.).

The reasons for the high proportion of temporary workers in Spain, despite the efforts of the afore mentioned reforms, have not been studied in depth. An exception is the paper by Dolado, García-Serrano, and Jimeno (2002). These authors claim that policy reforms did not have the desired effect on the share of temporary workers because of a simultaneous increase of temporary employment in the public sector.

An analysis of the Spanish manufacturing sector may provide an additional explanation to this puzzle. The Spanish manufacturing sector provides an interesting case study because it has a similar proportion of temporary employees as the aggregate economy and also shows a similar evolution in the period under study (see Figure A.1 in Appendix A). Interestingly, microeconomic data from the Survey of Managerial Strategies (Encuesta sobre Estrategias Empresariales, ESEE) shows that the proportion of temporary workers of domestic firms is 9 percentage points higher than the proportion of foreign-owned firms (where a firm is considered foreign in a given year if its proportion of foreign capital is larger than 50%). This suggests that the nationality of firm owners may influence the choice between temporary and permanent employment contracts.

There is a large literature analyzing the differences between domestic and foreign firms. The general conclusions of these studies are that foreign firms have higher labor productivity, a higher proportion of skilled workers and that they pay higher wages for workers of similar skills (Feliciano and Lipsey 1999, Conyon, Girma, Thompson, and Wright 2002, Griffith and Simpson 2003, Görg, Strobl, and Walsh 2007). This literature is an important precedent for this paper for two

reasons. First, it provides support for the idea that domestic and foreign firms may also differ in their hiring policies regarding temporary and permanent employment. Second, given that domestic and foreign firms have different characteristics, it is important to control for these characteristics when examining whether there is an effect of firm nationality on temporary employment.

The objective of this paper is to determine if, after controlling for a large set of observable firm characteristics and unobservable firm-specific time-invariant components, there is still a relationship between firm nationality and the type of employment contracts that firms offer. Identifying the characteristics of firms with a low proportion of permanent contracts is important, as it could help in the design of a labor reform aiming to provide the right incentives for firms to increase their use of permanent contracts.

For this purpose, I estimate standard censored Tobit and Heckman sample selection models (also known as type I Tobit and Heckman two-step models, respectively), using data from the Survey of Managerial Strategies (Encuesta sobre Estrategias Empresariales, ESEE), which includes a representative number of Spanish firms in the manufacturing sector during 1991-2005.

The estimations show that firm nationality has a significant effect over the probability that firms hire temporary workers and over the proportion of temporary workers for firms that choose to hire temporary workers. The size and significance of the effect depends on firm size. In the case of the Heckman estimations, for example, a higher proportion of foreign capital implies a lower average probability of hiring temporary workers for small and medium firms, but not for large firms. Likewise, a higher proportion of foreign capital implies a decrease in the average proportion of temporary workers (for firms who choose to hire temporary workers) for medium and large firms, but not for small firms.

In order to study the quantitative relevance of the effects, I calculate what the average probability of hiring temporary employees and the average proportion of temporary employees would be if firms changed their proportion of foreign capital to 50%. This exercise allows me to determine for which groups of firms a change in the proportion of foreign capital would have a significant effect on temporary employment. For example, in the case of the Heckman estimations, if firms with a proportion of foreign capital smaller than 50% were to change their proportion of foreign capital to 50%, the probability of hiring temporary employees would fall

in average 8.52 percentage points. However, the effect is not significant for firms with more than 50% of foreign capital changing their proportion of foreign capital to 50%. The analysis of Tobit estimations leads to similar conclusions.

I also find that the effect of foreign nationality on temporary employment is decreasing in firm size. In the case of the Heckman estimations, a change in the proportion of foreign capital of domestic firms to 50% implies a decrease in the probability of hiring temporary employees of 10.76 percentage points for small firms and of 5.28 percentage points for medium firms, and also implies a decrease in the proportion of temporary employees of 2.4 to 3.9 percentage points for medium firms and of 2.51 to 2.7 percentage points for large firms.

Fixed effects estimations show that there may be unobserved firm characteristics, like the managerial style or ability, which also influence the proportion of temporary employees. However, the estimated marginal effects of firm nationality are still statistically and quantitatively significant, which shows that firm nationality has an effect on the type of labor contracts offered, even after controlling for unobservable firm characteristics.

Given that the proposed econometric analysis considers a large number of covariates related with the structure of the firm, its production technology and demand, and the characteristics of its managers, as well as unobserved firm characteristics, the measured effects of firm nationality will not be caused by a correlation between firm nationality and other factors, and therefore, the coefficients of the reduced form analysis reflect a causal effect of firm nationality on temporary employment. These results are also robust to several alternative specifications.

This paper provides two contributions to the literature. First, I further the study of the differences between domestic and foreign firms. In particular, I show that domestic and foreign firms do not differ only in wages and productivity, but also on the types of labor contracts that they offer. Second, I provide a detailed study of the determinants of temporary employment in the Spanish manufacturing sector, focusing in particular on firm nationality and firm size. I find that foreign nationality has a negative effect on the probability of hiring temporary employees for small and medium firms, and on the proportion of temporary employees (for firms that hire temporary employees) of medium and large firms. These findings indicate that labor reform should provide different incentives to firms of different size.

The paper is organized as follows. Section 2 presents the theoretical framework. Section 3 describes the sample used in the estimations. Section 4 discusses the empirical model and estimation strategy. Section 5 presents the results. Finally, Section 6 concludes.

2. Theoretical framework

Firms determine their demand of permanent and temporary workers by solving a profit maximization problem. The demand of each firm will depend on the productivity, wages and hiring and dismissal costs of each kind of worker, and on the firm's need to adapt to fluctuations in the business cycle. All these factors may be firm specific, so they should be taken into account when attempting to determine whether firm nationality has an effect on the proportion of temporary workers. In what follows, I will explain how each of these factors affects the proportion of temporary workers and comment on the expected sign of this effect.

With respect to productivity, several papers show that permanent workers are in general more productive than temporary workers. For example, Sánchez and Toharia (2000) show that firm productivity diminishes as the proportion of temporary workers increases in the case of Spain. To control for this effect, I introduce variables related to productivity, human capital, on-the-job training and capital intensity, which are expected to have a negative effect on the proportion of temporary employees.

Firms may also differ in the wages they pay for each kind of worker. Unfortunately, the survey does not provide disaggregated information on wages, so I take an indirect approach and relate the wage mark-up (the ratio of wages of permanent and temporary workers) to firm specific variables. Wage determination models like efficiency wages, union bargaining and insider-outsider models show that in addition to external market forces, firm specific variables can also affect the wages of a given firm (see Katz 1986, Oswald 1985, Lindbeck and Snower 1988, respectively, for detailed surveys).

According to Nickell, Vainiomaki, and Wadhwani (1994), for example, the wage mark-up depends not only on internal factors, such as industry prices and productivity; external factors, such as the aggregate wage and unemployment rate; but also on product market conditions, represented by the firm's market share. The

latter effect arises because monopoly power generates rents, some of which can be expropriated by insiders (permanent workers) in the wage bargaining process.

As I have mentioned above, regressions will include variables to control for productivity differences. I will also include variables to account for the firm's market power. These variables are expected to have a negative effect on the proportion of temporary workers, through their effect on the wage mark-up (e.g. an increase in market share implies an increase in the wage of permanent workers relative to the wage of temporary workers, and thus an increase in the share of temporary employees).

Finally, adjustment costs models analyze the way in which firms adjust their employee staff to face economy-wide or industry-specific economic fluctuations (Nickell 1978, Sargent 1978, Kennan 1979, Hamermesh and Pfann 1996). Adjustment costs differ between permanent and temporary workers, and between increases and decreases in the number of employees. Temporary contracts have lower firing costs, so the proportion of temporary workers should decrease during recessions and increase in expansions.

Taking into account the literature analyzing the differences between firms of different nationalities, domestic and foreign firms may also differ in the way they react to changes in productivity, wages and economic fluctuations. For this reason, in Section 5.1 I present an alternative specification with interactions between firm nationality and the aforementioned controls.

Summarizing, regressions will include suitable controls to account for differences in productivity, wages and business cycle fluctuations. To the extent that I properly control for observable and unobservable firm characteristics, the estimated coefficients of firm nationality will measure the causal effect of nationality on the proportion of temporary employees.

3. Data description

3.1. The sample. The dataset comes from the Survey of Managerial Strategies (Encuesta sobre Estrategias Empresariales, ESEE)[2]. The sample is an unbalanced panel of 4,050 firms drawn from the Spanish manufacturing sector in the period 1990-2005.

The survey is national in scope and the sample is representative of the universe of manufacturing establishments of certain sizes. The sample has been selected across both activity sectors and size intervals, where size is determined by the number of workers. Two subpopulations have been distinguished, one formed by firms with more than 200 workers, and another by firms with 10 to 200 workers. For the first subpopulation the sample selection was exhaustive. For the second subpopulation, the sample was selected by random sampling across 20 activities and four employment size intervals.

Year 1990 is not considered because the measurement methodology of many variables changed in 1991. The sample has been cleaned by dropping observations of years in which the accounting period is incomplete, the export propensity is larger than 100 or the firm does not report some of the variables used in the estimations. Firms that were involved in a merger or acquisition are excluded from the sample. I also exclude firms that stop reporting because they become unreachable, disappear or stop cooperating. Finally, the selected sample includes all firms with at least two consecutive observations. The final sample has 2,136 firms and 16,075 observations. See Table B.1 in Appendix B for more details.

3.2. Variable definitions. In this section, I describe the variables used in the estimations.[3] The dependent variable is the proportion of temporary employees in total employment (pte). Total employment is the sum of temporary workers, full-time permanent workers and one half of part-time permanent workers. Temporary and full-time permanent workers are measured by the simple average of the quarterly number of employees when there has been significant variation, or

[2]This survey is funded by the Ministry of Industry of Spain and carried out by the Public Enterprise Foundation (Fundación Empresa Pública). See García, Jaumandreu, and Rodríguez (2002) for more information.

[3]All variables used in the estimations are directly obtained from the survey, except in the case of the logarithm of the production of goods and services in real terms, which was constructed using the production of goods and services (provided by the survey) and the IPRI (Industry Price Index), which is provided by the INE (National Institute of Statistics).

by the number of employees at the end of the year when the firm reports that this number has not changed much. Part-time permanent workers are measured by the number of workers at the end of the year.

There are two alternatives for the specification of the main explanatory variable: one is to use the proportion of foreign capital, and the other is to construct a dummy variable of firm nationality, which would be equal to 1 when the proportion of foreign capital is greater than 50% and equal to 0 otherwise. The main disadvantage of the dummy variable specification is that it does not provide enough temporal variation to apply fixed effects.[4] Therefore, I will use the continuous specification for the estimations. Nevertheless, I will keep the distinction between domestic and foreign firms for the section with descriptive statistics. For completeness, I will also present results using the dummy variable specification in Section 5.1.

To control for the effects of productivity and human capital, I introduce the average productivity of labor, the capital-to-labor ratio, the proportion of engineers and bachelors in total employment, and variables describing worker selection and worker formation. The average productivity of labor is given by the production of goods and services per worker and the capital-to-labor ratio is the net real capital in equipment goods divided by the total number of employees. Worker selection and worker formation are two categorical variables which can take on three values depending on whether the firm performs the activity itself, hires an external contractor to perform it, or does not practice the activity at all[5].

The wage mark-up will be mainly affected by productivity and the firm's market power. The variables used to describe productivity have already been discussed. The variables capturing the firm's market power are the market share, the market globalization index and the market atomization index. The market share is a weighted average of the market shares of all markets in which the firm operates, the market globalization index is the percentage of these markets that the firm identifies as foreign or domestic-foreign, and the market atomization index is the

[4]See the discussion in Section 3.3 for further details.

[5]Both worker selection and worker formation are reported only once every four years, but there is evidence that skill composition is relatively stable over time. Therefore, the combination of these variables with others describing productivity and human capital provides a good description of the skill composition of the firm.

percentage of markets served by the firm in which no firm has a significant market share and the firm itself does not have a market share larger than 10%.

The wage mark-up will also be affected by macroeconomic variables, such as the unemployment rate and the average aggregate wage. Given that these variables are the same for all firms at each period, their effect will be taken into account by including year dummies. At the same time, industry specific variables, like industry price, will be captured by industry dummies.

Several variables are included to control for the fluctuations in the demand of the firm, which may not always be in sync with the general business cycle. To capture the demand in the main market of the firm, I introduce two dummy variables: demand increase and demand decrease. These variables take the value 1 if the main cause of change in the price of the firm is an increase or decrease in demand, respectively, and 0 otherwise. Another control is market dynamism, which is a categorical variable based on a weighted market dynamism index, and takes into account all markets in which the firm participates. It indicates if the average demand is in a slump (index smaller than 35), is stable (index between 35 and 65), or is in a boom (index larger than 65). Finally, the estimations also include the logarithm of the production of goods and services deflated by the industry price index.

To complete the characterization of the firm, I include an additional set of variables including size, industry, location, age, age squared, legal type, type of good, outsourcing, export propensity, number of markets, financial situation, family business and public control. Moreover, given that time-industry effects may be particularly relevant in the Spanish case, estimations will also include interactions between time and industry dummies.

Firm size is determined by the number of employees, and three size intervals are considered: 10 to 50, 51 to 200 and more than 200 workers. With respect to industry, there are 20 activities, which correspond to the 3-digit CNAE classification. Location is determined by the region where the main industrial establishment is located. There are 17 regions that correspond to the Autonomous Communities of Spain.

Firms are classified according to their legal type, such as corporations and limited liability companies. The type of good that the firm produces may be a final good, an intermediate good or undefined. The variable representing outsourcing is

the ratio of subcontracted purchases over sales. The export propensity is measured by the percentage of exports over total sales and the financial situation of the firm is described by the annual cost of debt to financial institutions in the long-run and in the short-run.

To control for the possible effect of family business on the type of employment contracts offered, I include the proportion of owners and family of owners in the firm's management and administration. Finally, given that public firms may have other objectives besides profit maximization, I introduce a dummy variable of public control, which takes the value 1 if the firm has a share of public capital larger than 50% in any year in the examined period, and 0 otherwise. In an alternative exercise, I exclude all these firms from the sample.

3.3. Sample statistics. Table B.2 shows the distribution of firms according to ownership nationality, where a firm is considered domestic if its proportion of foreign capital is less than 50%, and is considered foreign otherwise.

Most observations represent domestic ownership (87.74% in comparison with 12.26% representing foreign ownership). Most of the firms in the sample (94.29%) never change nationality. Specifically, 1,801 firms have domestic nationality during the whole period, and 213 firms have foreign nationality during the whole period. Of the firms that changed nationality, 34.43% changed from domestic to foreign, 23.77% changed from foreign to domestic, and 41.80% changed nationality in more than one direction.

Table B.2 also shows descriptive statistics for the proportion of temporary employees (*pte*). The average proportion of temporary employees is 21.38% for the whole sample, but the average proportion is nearly 9 percentage points higher in the years in which firms are domestic (22.47%), in comparison with the years in which they are foreign (13.59%).

It is interesting to compare the differences in the proportion of temporary employees among firms that change nationality, given that with this exercise we are partly controlling for unobserved permanent heterogeneity. If we look at firms that change from foreign to domestic, for example, we can see that these firms have 9.16% of temporary workers in the years in which they were foreign, in comparison with 11.18% in years in which they were domestic. This small difference in the proportions of temporary workers is caused by the lack of within-firm variation in the sample when using the dummy variable of firm nationality, which already

suggests that fixed effect estimation is not going to work well with this specification, as was noted in the previous section. In the case of firms that change nationality from domestic to foreign, the difference in the proportions of temporary employees of years in which they were domestic or foreign is larger. However, the small number of observations of firms that change nationality in the sample still does not allow us to identify permanent unobserved heterogeneity when using the dummy variable specification for firm nationality.

A further analysis of the data shows there are important differences between foreign and domestic firms. Table B.3 shows the distribution of firms according to size. Domestic firms are in general smaller (70% have less than 50 workers) while foreign firms are larger (60.30% have more than 200 workers). The proportion of temporary workers is decreasing in size for domestic firms, but for foreign firms the relationship between size and *pte* is non-monotonic (medium firms have lower *pte* than small firms, but large firms have higher *pte* than medium firms). For each size, the proportion of temporary employees is larger for domestic firms than for foreign firms (differences are significant at 1%), but the difference diminishes as size increases (around 9% for small and medium firms and 3.26% for large firms).

Tables B.4 to B.12 in Appendix B present a further account of the differences between both types of firms. In general, the main activities of domestic and foreign firms are different. Domestic firms are mostly dedicated to textiles, metal products and food and tobacco, while foreign firms are mostly involved in electronic and electrical equipment, motor vehicles, and chemical products. Foreign firms are characterized by using more physical and human capital in their production process and are more capital intensive. Also, these firms invest more in the hiring and training of their workers, and prefer to subcontract these services. Foreign firms are more oriented to the international market, because they have a larger globalization index and their export propensity is 22.21 percentage points larger than that of domestic firms. In addition, foreign firms operate in markets that are more concentrated and their average market share is larger. There are also differences in the legal form of the firms. Almost all foreign firms are corporations, while domestic firms are more diversified in this aspect.

On the other hand, there are some characteristics in which the two types of firms are similar. For example, both kinds of firms prefer to be located in Catalonia or Madrid (38% of domestic firms and 55% of foreign firms), produce mainly

intermediate goods (50.34% of domestic firms and 67.87% of foreign firms) and prefer not to outsource the production of components or final goods. Finally, both kinds of firms operate in markets they consider relatively stable.

To summarize, there is clear evidence that foreign and domestic firms have different characteristics and behavior. It is therefore necessary to control for all these characteristics when investigating whether firm nationality is an important factor in determining the type of labor contract that firms offer.

4. Empirical model and estimation strategy

The objective of the empirical analysis is to determine if there is a relationship between firm ownership nationality and the share of temporary employees, even after controlling for observed and unobserved firm characteristics.

When choosing the econometric model, it is necessary to take into account that 22.9% of the observations correspond to firms with a zero proportion of temporary contracts (corner solutions). One possibility is to apply a censored regression model like the standard censored Tobit model (Tobin 1958), also known as type I Tobit model.[6]

The standard censored Tobit model that allows for lower censoring assumes that the *observed* dependent variable, y_{jt}, satisfies

$$y_{jt} = \max(0, y^*_{jt}), \tag{1}$$

where $j = 1, \ldots, N$ is the firm index and $t = 1, \ldots, T$ is the period index. y^*_{jt} is the latent variable generated by the classical linear regression model

$$y^*_{jt} = x'_{jt}\beta + \mu_{jt}, \tag{2}$$

where x_{jt} is a vector of regressors, including 1 for the intercept, and β is the corresponding vector of parameters. The errors μ_{jt} are independent and follow a distribution $N(0, \sigma^2_\mu)$, conditional on x_{jt}. Note that x_{jt} is independent of μ_{jt} but the relation between x_{jt} and μ_{js}, for $s \neq t$ is unspecified.

[6]Strictly speaking, the dependent variable has potentially both lower and upper censoring (it can only take on values in the unit interval). However, the upper censoring does not take place in practice, so I will use a model which only allows for lower censoring. It is also important to remark that, even though the range of the dependent variable is bounded, the estimated model is still a valid linear approximation of the function determining *pte* as a function of the explanatory variables.

The main limitation of this model is that the same set of variables and coefficients determine both the probability that an observation is censored and the value of the dependent variable. In our case, this means that all explanatory variables affect the probability of entry into the temporary employment market and the proportion of temporary contracts offered in the same way.

This shortcoming can be overcome by using a more flexible estimation approach, such as the sample selection model of Heckman (1979), also known as Heckman two-step model. In this model, a different set of variables and coefficients determines the probability of a zero outcome and the value of the dependent variable given that it is not zero. In particular, it could be possible that nationality affects the proportion of temporary contracts offered (given that the proportion is positive), but not the probability of entry into the market of temporary employees.[7]

Heckman's sample selection model is characterized by the following two equations:

$$y_{2jt} = 1[x'_{jt}\,\alpha + v_{jt} \geq 0], \tag{3}$$

$$y^*_{1jt} = x'_{1jt}\,\delta + u_{jt}. \tag{4}$$

The first equation is the selection or entry equation and the second one is the outcome or level-of-use equation. y_{2jt} is a binary entry indicator which takes the value 1 if $x'_{jt}\,\alpha + v_{jt} \geq 0$ and 0 otherwise. x_{jt} is a k-vector of regressors and x_{1jt} is an m-vector of regressors with $m \leq k$.

The entry equation determines the probability of entry into the market of temporary contracts and the level-of-use equation determines the proportion of temporary contracts offered, given that the firm has decided to enter into this market.

The underlying assumptions of this model are: (i) $(x_{1jt}, x_{jt}, y_{2jt})$ are always observed, but y^*_{1jt} is observed only when $y_{2jt} = 1$, (ii) u_{jt}, and v_{jt} are independent of x_{jt} and x_{1jt}, (iii) $v_{jt} \sim N(0,1)$, and (iv) $E(u_{jt}|v_{jt}) = \gamma\, v_{jt}$. The model is estimated through a standard two-step procedure, which gives consistent estimators under the above assumptions.

[7]Even though the Tobit model places additional restrictions to the model equations, in comparison with the Heckman model, it is interesting to compare the results of both estimations, in order to see the effect of these additional restrictions on the estimation results. For this reason, and also for completeness, I will estimate both Tobit and Heckman models for the different specifications in the paper.

In Section 5 I present the estimations of the Tobit and Heckman models. The dependent variable of interest in both models is pte_{jt}. The regressors included in x_{jt} are interactions of the proportion of foreign capital with each category of size (these are the main explanatory variables), dummy variables of size, legal form, type of good, market dynamism, demand increase and demand decrease, public participation, worker selection, worker formation, industry, location, and year (plus interactions between time and industry dummies); and continuous variables like the capital-labor ratio, age, age squared, proportion of engineers and bachelors, average productivity of labor, market share, market globalization index, market atomization index, production of goods and services in logs, outsourcing, export propensity, number of markets, cost of debt in the long-run, cost of debt in the short-run, and family control.

In the case of Heckman's sample selection model, it is well known that if the set of explanatory variables is the same in the selection and outcome equations, the coefficients are identified only due to the non-linearity of the inverse Mill's ratio. However, if $x'_{jt}\alpha$ does not have much variation in the sample, the inverse Mill's ratio will be approximately linear. Therefore, in order to guarantee the identification of the parameters, it is better to have non-trivial exclusion restrictions.

Consequently, the regressors included in the x_{1jt} vector will be the same than in x_{jt}, with the exception of the dummies of worker selection, the outsourcing indicator and family control. Moreover, in order to have additional exclusion variables and with the aim of picking up the effect of the legal reforms carried out in 1994, 1997, and 2001, I will include dummy variables for the labor reforms, where each of these dummies is one for the year of the reform and the following years until a new reform, and 0 otherwise. Since the objective of these reforms was to discourage the use of fixed-term contracts, they will mainly affect the decision of hiring temporary workers or not.

4.1. Fixed effects estimation. Until now I have considered that the proportion of temporary contracts is affected by a large set of observable variables. However, it is also reasonable to think there is a firm-specific time-invariant unobserved heterogeneity component (η_j), which could be related with the managerial style or ability of the firm. For example, firms with higher managerial skills will tend to hire more permanent workers, just like firms with higher productivity. Therefore,

if foreign firms have higher (lower) ability, the effect of foreign nationality on the proportion of temporary workers will be overestimated (underestimated).

This endogeneity problem can be solved by applying fixed effects techniques,[8] which means that no restrictive assumptions on the distribution of η_j are imposed.

In the Heckman model, assuming that the unobserved component is only in the outcome equation, it is possible to obtain fixed-T consistent estimators by applying a two-step estimation procedure, and estimating in the second step a transformed model in deviations with respect to individual means. In the case of the entry equation of the Heckman model (which is a Probit model) or the Tobit model, on the other hand, it is not possible to obtain fixed-T consistency without assuming a particular distribution for the unobserved effect[9]. Moreover, I consider the selection equation of the Heckman model as a reduced form equation, which is included in the estimations to take into account the selection process, and thus obtain better estimates of the level of use equation. In other words, I am not interested in determining the causal effect of the independent variables on the probability of selection, but in correcting the estimates of the level equation. For these reasons, I will only consider fixed effects in the level of use equation of the Heckman model.

The empirical model to be estimated is:

$$y_{2jt} = 1[x'_{jt}\alpha + v_{jt} \geq 0], \tag{5}$$

$$y_{1jt} = x'_{1jt}\delta + \gamma\hat{\lambda}_{jt} + \eta_j + u_{jt}. \tag{6}$$

As before, y_{2jt} is the binary entry indicator and v_{jt} and u_{jt} are idiosyncratic disturbances, independent of x_{jt} and x_{1jt}. η_j is a time-invariant individual-specific fixed effect.

I estimate this model in two steps. The first step consists of a Probit estimation of (5). The second step consists of OLS estimation of (6), including as a regressor the estimated inverse Mills ratio, $\hat{\lambda}_{jt} \equiv \lambda(x'_{jt}\hat{\alpha})$, obtained from the first

[8]It is also possible to apply a correlated random effects approach, but this is more restrictive because it is necessary to assume some distributional form of the unobserved component to relate it with the covariates.

[9]In general, is not possible to obtain fixed-T consistency in non-linear models, like Probit or Tobit. There are some exceptions, like the conditional Maximum Likelihood for the static panel Logit model (Andersen 1970), or Honoré and Kyriazidou's (2000) estimator for the dynamic panel Logit model, but these models impose restrictive assumptions which hold only in certain specific cases.

step. To account for unobserved heterogeneity, the second equation is estimated in deviations from individual means.

Finally, the variables included in x_{jt} will be the same variables as in the case without fixed effects, with the exception of the dummies of public control, location and industry, given that they do not have enough within-firm variation in the sample, and thus cannot be separately identified from permanent unobserved heterogeneity.

5. Results

Table 1 shows the average marginal effects of the baseline specification for the type I Tobit and Heckman sample selection models. In the Tobit case, I report the average marginal effects on the expected value of the proportion of temporary workers (pte), given that this value is greater than zero, $E(pte \mid x,\, pte > 0)$. In the Heckman case, I report the average marginal effects on the probability of entering the market of temporary contracts, $Pr(pte > 0 \mid x)$, and the expected value of pte given that it is greater than zero, $E(pte \mid x,\, pte > 0)$.

[TABLE 1 ABOUT HERE.]

The dependent variables in all equations are expressed in percentage terms. Therefore, the marginal effects measure the percentage change in the proportion of temporary workers (Tobit and level-of-use equations) or the probability of entry in the market for temporary workers (entry equations).

It is worth noticing that the first three entries in Table 1 are not the average interaction effects between pfk and size, but the average marginal effects of the proportion of foreign capital (pfk) calculated for groups of firms of different size. The fourth entry is the average marginal effect of pfk considering all firms.

Given that the proportion of foreign capital is a continuous variable, the marginal effects shown in Table 1 show the effects of an infinitesimal change in the proportion of foreign capital. These average marginal effects already show the statistical significance of the effect of changes in the proportion of foreign capital on temporary employment. However, it is also important to have an idea of the magnitude of these effects.

Table 2 shows the average effect of a change in the proportion of foreign capital to 50%. For firms with less than 50% of foreign capital (domestic firms), this exercise simulates what would happen if they changed their proportion of foreign capital to the minimum possible amount to be considered foreign (according to our definition of nationality). Likewise, for firms with more than 50% of foreign capital (foreign firms), the exercise simulates what would happen if they changed their proportion of foreign capital to the minimum possible amount to be considered domestic.[10]

[TABLE 2 ABOUT HERE.]

In the Tobit model the marginal effects of the proportion of foreign capital are negative and significant at the 1% level for the three sizes. The average effect for all firms is also significant. With respect to the average effect of changing the proportion of foreign capital to 50%, it is significant for firms changing nationality from domestic to foreign, but not for firms changing from foreign to domestic. Specifically, if all domestic firms would change their proportion of foreign capital to 50%, the proportion of temporary employees would fall in average 2.9 percentage points. Moreover, the effect is larger for small firms and medium firms changing from domestic to foreign.

In the level of use equation of the Heckman model without fixed effects, the estimated effects of pfk are negative and significant at the 1% level for medium and large firms, but are not significant for small firms, or when considering all firms.

If we look at the effect of changing pfk to 50%, we can see that, as in the Tobit case, the effect is only significant for firms changing their nationality from domestic to foreign. Specifically, if medium sized domestic firms would change their proportion of foreign capital to 50%, the proportion of temporary employees would fall in average 3.9 percentage points, and if large domestic firms would

[10]Of course, it could well be the case that a national or domestic shareholder gains control of the firm with less than 50% of the firm's capital. However, in this exercise, the 50% threshold is only used to test the quantitative importance of a discrete change in the proportion of foreign capital. Considering a different threshold would not have a significant impact on the sign or statistical significance of the effects, but would affect the magnitude of such effects.

change pfk to 50%, the proportion of temporary employees would fall in average 2.7 percentage points.

With respect to the entry equation, the effect of pfk on the probability of entry into the market of temporary workers is negative considering all firms, and also for small and medium firms. A change of pfk to 50% has a greater impact for firms going from domestic to foreign than for firms going from foreign to domestic. Specifically, if all domestic change their pfk to 50%, the probability of hiring temporary employees would fall in average 8.53 percentage points. The effect for foreign firms gaining domestic nationality is significant only for small firms: if all small foreign firms reduce pfk to 50%, the probability of hiring temporary employees would increase 4.99 percentage points in average.

The difference in significance between the marginal effects of the entry and level of use equations in the Heckman model indicate the importance of allowing different processes to determine the probability of entry into the market for temporary workers, and the share of temporary workers, when the firm decides to hire temporary workers.

For example, in the Tobit model, a higher proportion of foreign capital for small firms increases the probability of hiring temporary employees, but also increases the share of temporary employees for firms that decide to hire temporary workers (which is due to the fact that the same set of coefficients determines both quantities). In the Heckman model, on the other hand, a higher proportion of foreign capital for small firms affects the probability of hiring temporary employees, but not the share of temporary workers given that the firm chooses to hire temporary workers.

In the Heckman model without fixed effects, the coefficient of the inverse Mill's ratio is negative and significant, which means that variables that increase (decrease) the probability of entering the market of temporary employees have an indirect negative (positive) effect on the level of temporary workers.

Fixed effects estimations reinforce the results of the estimations without fixed effects. Marginal effects maintain sign and significance in the entry and outcome equations, which means that firm nationality has an effect on the proportion of temporary employees even after controlling for time invariant unobserved effects.

Comparing the results of the Heckman estimations with and without fixed effects, we can see that the magnitudes of the effects of firm nationality are similar

in both estimations, although they are slightly larger for the model without fixed effects. For example, for medium sized domestic firms changing pfk to 50%, the effect is 2.38 percentage points for the fixed effects estimation, and is 3.9 percentage points for the model without fixed effects.

The coefficient of the inverse Mill's ratio, which measures the extent of sample selection, becomes non significant once we introduce fixed effects. This means that the correlation we had found between the errors in the entry and level equations in the first Heckman estimation may be due to unobserved characteristics. Once we take into account these unobserved differences, the selection process looses significance. This finding shows the importance of including fixed effects into the analysis.

The above analysis allows us to conclude that, after controlling for a large set of observable firm characteristics and also for unobservable heterogeneity across firms, there is a significant effect of the proportion of foreign capital on the type of employment contracts offered by firms.

With respect to the variables used as controls, in general the marginal effects have the expected sign. The marginal effects of the variables related with productivity and human capital (proportion of engineers and bachelors, average productivity of labor and capital-to-labor ratio) are negative and generally significant, which confirms our previous assertion that more productive firms tend to have a smaller proportion of temporary workers.

A higher market share implies a decrease in the probability of hiring temporary employees, and also on the proportion of temporary employees (except in the fixed effects estimation). Therefore, the marginal effect of market share has the expected sign. The market globalization index is significant only in the entry equation of the Heckman estimations: firms that are more globalized have a lower probability of hiring temporary employees. With respect to the market atomization index, it does not have a clear effect on pte, since sign and significance depend on the estimation.

The effects of the dummies of market dynamism are significant. Firms hire more temporary workers in expansions, and hire less temporary workers in recessions, as expected. The effect of recessions is larger in absolute value than the effect of expansions in the level-of-use equations of the Heckman estimations, which means that firms adjust their number of temporary workers more during

recessions than during expansions. With respect to the demand in the main market, a demand decrease has a negative effect on *pte* (except in the fixed effects estimation), but it is interesting to notice that the effect is also negative for an increase in demand.

Therefore, it is clear that in slumps the need for reducing costs drives firms to dispense with temporary employees. However, during booms firms may have incentives to hire either permanent or temporary workers. It is true that firms prefer to hire temporary employees to avoid high dismissal costs in the future, but if the firm believes that the expansion will last for a long time the lower probability of firing workers will encourage the hiring of permanent workers. Thus, the adjustment of labor demand depends on the phase of the cycle and on the firm's expectations of the duration of the boom or slump.

Another variable with a strong effect is the dummy of public capital. In the case of the Heckman model without fixed effects, for example, having a share of public capital larger than 50% in at least one year, implies an average decrease of 11.74 percentage points in the proportion of temporary employees.

Finally, it is interesting to analyze the effects of the labor reforms in the entry equation. These marginal effects are calculated comparing the years after the reform (until a new reform) with the years corresponding to the previous reform. Results show that the reform corresponding to 1994 had a positive effect in the share of temporary workers, which means that this reform failed to discourage the use of fixed-term contracts. On the other hand, the reforms of 1997 (in comparison with that of 1994) and 2001 (in comparison with 1997) had a negative effect on the probability of entering the temporary contracts market, which is consistent with the objective of those reforms.

In conclusion, the estimations of type I Tobit and Heckman sample selection models show that nationality has a negative effect on the probability of hiring temporary employees or the proportion of temporary employees, depending on firm size. These findings support the hypothesis that there is a causal relationship between firm ownership nationality and the share of temporary employees, after accounting for observable and unobservable firm characteristics.

5.1. Robustness checks. I have estimated alternative specifications to test the robustness of the results of the baseline specification. First, taking into account that there are significant differences in the characteristics of foreign and domestic

firms, I include a set of interactions between the proportion of foreign capital and several control variables (all control variables except public control, and time, industry and region dummies, for which there is not enough variation to accurately estimate the coefficients corresponding to the interactions). The results of this estimation (Specification II) are presented in Table 3.

[TABLE 3 ABOUT HERE.]

[TABLE 4 ABOUT HERE.]

Looking at Table 3 we can see that the results of Specification II are similar to those of the baseline specification, although there are some differences. For example, in the Heckman specification with fixed effects, the marginal effect of pfk remains significant for medium firms, but becomes non significant for large firms. This loss of significance may be due to the fact that many variables do not have enough variation in the sample, which makes it difficult to accurately estimate the coefficients of a large number of interactions in the fixed effects specification.

Nevertheless, Table 3 shows there is a significant effect of pfk on the proportion of temporary workers, which is larger for medium firms in comparison with small and large firms. Table 4 shows the effect of a change in pfk to 50%, and provides additional support to the results of Table 3.

In order to see the effect of the interactions, Table 5 shows the effect of pfk for different groups of firms. For example, the marginal effects of the level of use equation of the Heckman model without fixed effects shows that the effect of foreign nationality is smaller for firms that are more capital intensive, are older, have higher labor productivity, and have a higher market globalization index. The level of use equation of the Heckman model with fixed effects shows for most groups the effect of nationality is non significant, which is due to the loss of significance of this variable in the fixed effects specification, as discussed above.

[TABLE 5 ABOUT HERE.]

Second, given that public firms may have other objectives besides profit maximization, all firms with a positive proportion of public capital in some year are excluded from the sample (Specification III). These firms represent only a 2.4% of the firms in the sample (51 firms and 440 observations are dropped). Table 6 shows the marginal effects of pfk and Table 7 shows the average effects of a change in pfk to 50%. Comparing these results with those of the baseline estimation, we can see that when firms with public participation are not considered, the conclusions of the analysis do not change significantly. In the Heckman case, for example, the effects are slightly larger for the entry equation and slightly smaller for the level of use equation, but they keep sign and significance (except in the case of the effect of a change in pfk to 50% for large firms in the Heckman model without fixed effects).

[TABLE 6 ABOUT HERE.]

[TABLE 7 ABOUT HERE.]

Third, firms' adaptation to fluctuations in demand or technology may not be instantaneous. Specification IV includes the lags of the dummies of demand increase and decrease and the lag of the production of good and services, instead of their contemporaneous values. Table 8 shows the marginal effects of pfk and the lagged variables, and Table 9 shows the effects of a change of pkf to 50%. The main difference with the baseline specification is that in the level equation of the model with fixed effects, the marginal effects of foreign capital loose significance. In particular, the effect is no longer significant for medium firms, and is significant for large firms at a 10% level. Moreover, the effects of a change in pfk to 50% in the level equation are no longer significant for medium and large firms. For the other models, the results are similar to those of the baseline specification. With respect to the marginal effects of the lagged variables, they are also less significant than their contemporaneous counterparts in the baseline specification. Demand increase is in general positive but non significant, and demand decease is negative but smaller and in some cases non significant. The effect of changes in the production of goods and services continues to be positive and significant, but

it is smaller than the effect of the contemporaneous counterpart in the baseline specification.

[TABLE 8 ABOUT HERE.]

[TABLE 9 ABOUT HERE.]

Finally, it is interesting to test the robustness of the results to the definition of firm nationality. Specification V replaces the continuous specification of firm nationality (proportion of foreign capital) by a dummy variable which is equal to 1 if the firm has a share of foreign capital greater than 50%, and is equal to 0 otherwise. Given that the dummy variable does not have much variation in the sample, this specification is not amenable to the inclusion of fixed effects. Table 10 reports estimation results for this specification. We can see that this specification leads to similar results as the baseline specification. In the entry equation of the Heckman model, for example, foreign nationality has an effect for small and medium firms, but not for large firms. In the level equation of the Heckman model, foreign nationality has an effect for medium and large firms, but not for small firms. Both results correspond to the findings of the benchmark model. As for the magnitude of these effects, having a proportion of foreign capital higher than 50% implies an average decrease of 5.47 percentage points in the share of temporary employees for medium firms, and a decrease of 2.79 percentage points in the share of temporary employees for large firms.

[TABLE 10 ABOUT HERE.]

Summarizing, the above analysis shows that the results of the baseline specification are robust to the introduction of interactions and lags of relevant control variables, the exclusion of firms with public participation, and the use of alternative definitions of firm nationality.

6. Conclusion

In this paper, I present a study of the effects of firm nationality on the share of temporary employees focusing on the Spanish manufacturing sector. For that purpose, I estimated pooled type I Tobit and Heckman sample selection models, and fixed effects Heckman sample selection models, regressing the proportion of temporary employees on the proportion of foreign capital and a large set of observable firm-specific variables. The sample used in the estimations comes from the Survey on Managerial Strategies (Encuesta Sobre Estrategias Empresariales, ESEE) in the period 1991-2005.

The main finding of the paper is that, after controlling for observable and unobservable firm characteristics, there is a significant relationship between the proportion of foreign capital and the employment contracts that firms offer. Therefore, there is evidence in favor of the hypothesis that the lower proportion of temporary contracts observed in firms with a high proportion of foreign capital is not caused by a composition effect (firm or industry composition), but that there is a causal effect of firm nationality.

In order to study the quantitative relevance of the effects, I calculate what the average probability of hiring temporary employees and the average proportion of temporary employees would be if firms changed their proportion of foreign capital to 50%. This exercise allows me to determine for which groups of firms a change in the proportion of foreign capital would have a significant effect on temporary employment. For example, in the case of the Heckman estimations, if firms with a proportion of foreign capital smaller than 50% were to change their proportion of foreign capital to 50%, the probability of hiring temporary employees would fall in average 8.52 percentage points. However, the effect is not significant for firms with more than 50% of foreign capital changing their proportion of foreign capital to 50%. The analysis of Tobit estimations leads to similar conclusions.

I also find that the effect of foreign nationality on temporary employment is decreasing in firm size. In the case of the Heckman estimations, a change in the proportion of foreign capital of domestic firms to 50% implies a decrease in the probability of hiring temporary employees of 10.76 percentage points for small firms and of 5.28 percentage points for medium firms, and also implies a decrease in the proportion of temporary employees of 2.4 to 3.9 percentage points for medium firms and of 2.51 to 2.7 percentage points for large firms.

The comparison of the Tobit and Heckman estimations show the importance of allowing two different processes to determine the probability of having temporary employees, and the proportion of temporary employees, given that the firm decides to employ some of them.

Fixed effects estimations show that there may be unobserved firm characteristics, like the managerial style or ability of the firm, which also influence the proportion of temporary employees. However, the estimated marginal effects of firm nationality are still statistically and quantitatively significant, which shows that firm nationality has an effect on the type of labor contracts offered, even after controlling for observable and unobservable firm characteristics.

The difference in the incentives to use permanent contracts between foreign and domestic firms may be due to differences in the degree of risk aversion, the intertemporal discount rate or the time horizon taken into account to make decisions, which may be affected by the nationality of the firm. It may also be the case that foreign firms have directives from parent firms in the home country concerning the hiring and training of workers, and these directives may deprecate the use of temporary contracts.

Another reason may be that foreign firms have higher bargaining power, and can therefore negotiate more favorable terms of permanent contracts with the government or unions of the host country. Larger domestic firms may also benefit from this additional bargaining power, and this may be the reason why the effect of nationality is smaller for large firms.

The results raise interesting questions about why domestic firms prefer the flexibility of temporary contracts and foreign firms the greater productivity or experience of permanent contracts. Addressing this question would complement the results of the present paper, and is an interesting direction for further research.

Appendix A: Labor market indicators

TABLE A.1. Strictness of Employment Protection Legislation (EPL)

Country	Regular Contracts			Fixed-Term Contracts		
	1990	1998	2003	1990	1998	2003
Australia	1	1.5	1.5	1.3	1.3	1.3
Austria	2.6	2.9	2.4	1.8	1.8	1.8
Belgium	1.5	1.7	1.7	5.3	1.5	1.5
Canada	0.9	1.3	1.3	0	0	0
Denmark	1.6	1.5	1.5	1.3	2.3	2.3
Finland	2.7	2.3	2.2	3.3	3.3	3.3
France	2.3	2.3	2.5	3.5	4	4
Germany	2.7	2.7	2.7	3.5	1.8	1.8
Greece	2.5	2.3	2.4	4	4	4.5
Ireland	1.6	1.6	1.6	0	0	0.8
Italy	2.8	1.8	1.8	5.3	4	2.5
Japan	2.7	2.4	2.4	1	0.5	0.5
Korea		2.4	2.4		0.8	0.8
Netherlands	3.1	3.1	3.1	1.5	0.8	0.8
New Zealand		1.4	1.7		0.3	1.5
Norway	2.4	2.3	2.3	3.3	3.3	3.3
Poland		2.2	2.2		1	0
Portugal	4.8	4.3	4.2	2.3	2.3	1.8
Spain	**3.9**	**2.6**	**2.6**	**1.5**	**2.5**	**3**
Sweden	2.8	2.9	2.9	2.7	1.8	1.8
Switzerland	1.2	1.2	1.2	1.3	1.3	1.3
Turkey		2.6	2.6		4.3	4.3
United Kingdom	0.8	0.9	1.1	0	0	0.3
United States	0.2	0.2	0.2	0	0	0

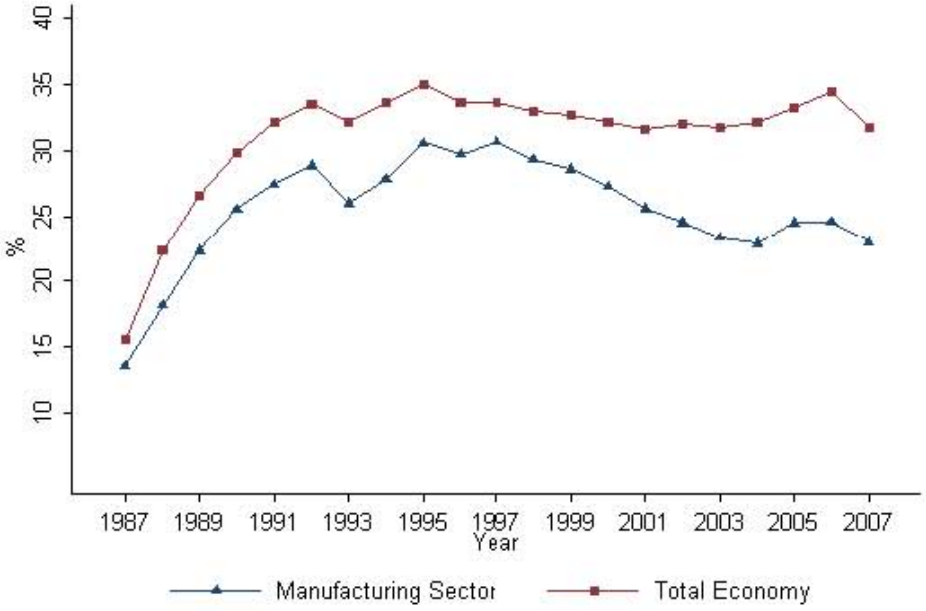
Indices range from 0 (least stringent) to 6 (most stringent).
Data Source: OECD.Stat.

TABLE A.2. Labor Indicators

Country	Average Employment Rate			Average Unemployment Rate			Average Share of Temporary Employment		
	1980/9	1990/9	2000/06	1980/9	1990/9	2000/07	1980/9	1990/9	2000/06
Australia	65.67	67.87	71.55	7.58	8.81	5.87		4.58	4.78
Austria	63.85	68.73	69.21	3.30	3.94	4.31		6.97	8.08
Belgium	53.17	56.74	60.56	10.83	8.44	7.30	5.90	6.11	8.66
Canada	67.76	69.09	73.12	9.40	9.55	7.08		11.71	12.80
Czech Republic		69.42	65.81		5.21	7.90		8.16	9.11
Denmark	76.01	75.44	77.09	7.60	7.46	4.61	11.43	11.10	9.63
Finland	72.72	65.00	68.27	4.89	11.88	8.86		17.64	16.37
France	61.53	59.78	62.43	8.90	11.10	9.58	5.87	11.91	13.50
Germany	63.58	65.58	66.39	6.05	7.77	9.30	10.86	11.03	12.83
Greece	57.04	56.19	59.73	7.65	9.11	9.88	18.00	12.02	12.26
Hungary		54.71	56.97		9.55	6.37		6.44	7.15
Iceland	78.02	84.67	87.46		3.85	2.83		12.45	10.56
Ireland	52.51	55.89	67.07	15.32	12.16	4.22	7.81	8.57	4.06
Italy	53.61	52.98	57.18	10.08	11.26	8.60	5.48	7.25	10.91
Japan	70.71	74.27	74.58	2.50	3.04	4.80	10.09	10.71	13.39
Luxembourg	59.61	60.55	63.14	2.48	2.48	3.34	3.83	2.94	4.21
Mexico		61.66	62.51		3.98	3.04		21.48	20.22
Netherlands	54.41	65.95	72.67	10.13	5.96	3.97	7.99	10.48	14.73
Norway	77.46	75.97	77.83	2.74	4.79	3.94		11.45	9.63
Portugal	67.28	69.14	72.43	7.14	5.63	5.91	16.98	13.39	20.36
Slovak Republic		60.50	57.65		13.16	17.41		3.92	5.10
Spain	**49.93**	**50.96**	**61.62**	**17.43**	**19.60**	**10.80**	**21.51**	**32.92**	**32.46**
Sweden	81.69	75.67	75.65	2.76	7.41	6.15		15.35	15.30
Switzerland	76.71	79.62	80.12	0.63	3.32	3.56		12.34	12.40
Turkey	56.73	54.00	48.21	8.52	7.79	9.46	17.58	19.03	15.87
United Kingdom	69.35	71.35	73.85	10.17	8.04	4.96	6.20	6.37	6.03
United States	70.76	74.53	74.66	7.27	5.76	5.10		4.72	4.11
Europe	60.34	61.24	62.49	9.05	9.62	8.79	8.21	11.74	13.67
OECD countries	65.70	66.66	67.55	7.13	7.05	6.54	9.15	11.04	13.05

Data Source: OECD.Stat

FIGURE A.1. Proportion of temporary employees



Appendix B: Description of the sample

TABLE B.1. Reasons for deletion of observations

Reason	Obs.	Firms
Accounting period is incomplete	45	
Propensity to export is greater than 100	24	
Missing variable	3,871	
Firm is involved in a process of merger, acquisition or scission	5,271	
Firm is unreachable, has disappeared or does not cooperate	34,134	
Firm with less than two consecutive observations	1,330	
Total	44,675	
Initial Sample	60,750	4,050
Deleted observations	44,675	1,914
Final sample	16,075	2,136

TABLE B.2. Firm Distribution in the Sample

	Observations		Firms		Prop. of Temp. Employees			
	Num.	%	Num.	%	Mean	St.dev.	Min	Max
Nationality								
Domestic	14,105	87.74	-	-	22.47	24.16	0	100.00
Foreign	1,970	12.26	-	-	13.59	16.08	0	94.29
Total	**16,075**	**100.00**	**2,136**	**100.00**	**21.38**	**23.50**	**0**	**100.00**
Never change								
Domestic	13,631	84.80	1,801	84.32	22.58	24.21	0	100.00
Foreign	1,412	8.78	213	9.97	12.82	14.88	0	94.29
Total	**15,043**	**93.58**	**2,014**	**94.29**	**21.67**	**23.67**	**0**	**100.00**
Change								
Foreign to domestic								
Domestic	112	0.70	-	-	11.18	9.41	0	37.89
Foreign	149	0.93	-	-	9.16	10.45	0	60.42
Total	261	1.62	29	1.36	10.04	10.04	0	60.42
Domestic to foreign								
Domestic	128	0.80	-	-	23.14	27.20	0	96.64
Foreign	173	1.08	-	-	16.74	21.18	0	92.19
Total	301	1.87	42	1.97	19.46	24.09	0	96.64
Several changes								
Total	470	2.92	51	2.39	19.27	21.02	0	95.86
Total firms that change	**1,032**	**6.42**	**122**	**5.71**	**17.27**	**20.54**	**0**	**96.64**

Data Source: ESSE

TABLE B.3. Firm Size

Size	Domestic		Foreign		Differences
(number of workers)	Obs. (%)	Mean Pte	Obs. (%)	Mean Pte	in Pte
1: [10, 50]	70.00	23.70	10.05	14.64	9.06***
		[0.25]		[1.73]	[1.75]
2: [51, 200]	14.90	22.20	29.64	12.84	9.36***
		[0.51]		[0.69]	[0.86]
3: $>$ 200	15.10	17.05	60.30	13.79	3.26***
		[0.40]		[0.40]	[0.57]
Total	14,105		1,970		

Data Source: ESSE

Standard errors in brackets; *** significant at 1%.

TABLE B.4. Firm Activity

Activity	Domestic		Foreign	
	Obs. (%)	Mean Pte	Obs. (%)	Mean Pte
1: Meat processing	3.44	26.55	1.73	18.25
2: Food and tobacco	10.22	29.08	7.31	23.35
3: Beverages	1.96	15.58	1.62	22.34
4: Textiles and clothing	12.19	23.43	5.94	11.16
5: Leather and footwear	4.47	35.01	0.1	0.14
6: Wood product manufacturing	3.57	27.85	1.07	16.35
7: Paper manufacturing	2.59	12.85	4.21	8.06
8: Publishing, printing and reproduction of recorded media	6.46	13.22	1.52	10.60
9: Chemical products	4.74	11.83	10.86	7.83
10: Rubber and plastic products	4.57	24.42	8.22	18.77
11: Non-metallic mineral products	7.07	23.97	4.06	13.05
12: Ferrous and nonferrous metals	2.55	17.81	2.49	12.52
13: Metal products	10.06	24.10	9.49	11.49
14: Industrial and agricultural machinery	6.62	15.49	9.14	9.07
15: Office machinery and computers	1.32	16.24	2.84	19.17
16: Electrical and electronic machinery and equipment	4.18	21.79	11.93	15.35
17: Motor vehicles	3.13	20.25	12.13	13.17
18: Other transport equipment	1.79	19.64	2.44	13.02
19: Furniture manufacturing	6.44	26.29	1.22	20.76
20: Other manufacturing	2.64	22.82	1.68	10.26
Total	14,105		1,970	

Data Source: ESSE

TABLE B.5. Descriptive Statistics of Continuous Variables

Variable	Nat.	Mean	St.dev.	Min	Max
Proportion of engineers and bachelors (%)	Domestic	3.00	5.57	0	71.4
	Foreign	6.64	7.57	0	49.9
Proportion of owners in firm's management and admin. (%)	Domestic	3.98	5.87	0	100
	Foreign	0.09	0.63	0	12.5
Export propensity (%)	Domestic	12.45	22.24	0	100
	Foreign	34.66	29.72	0	100
Market globalization index (%)	Domestic	14.79	33.58	0	100
	Foreign	36.53	44.78	0	100
Capital-to-labor ratio	Domestic	36,985	54,762	5.73	984,279
	Foreign	87,576	110,035	714.20	1,307,827
Production of goods and services (in logs)	Domestic	14.91	1.67	9.28	21.20
	Foreign	17.36	1.25	12.97	22.69
Average productivity of labor	Domestic	97,551	94,594	533	1,584,209
	Foreign	189,433	209,396	20,335	2,452,482
Firm age	Domestic	21.44	19.77	0	224
	Foreign	30.62	21.91	0	128
Market share (%)	Domestic	8.48	16.67	0	100
	Foreign	19.95	21.64	0	100
Market atomization index (%)	Domestic	23.92	40.86	0	100
	Foreign	10.41	28.20	0	100
Subcontracted purchases over sales (%)	Domestic	8.96	17.86	0	100
	Foreign	9.23	18.68	0	100
Annual cost of debt to financial institutions in the long-run (%)	Domestic	1.72	3.63	0	31
	Foreign	1.16	2.96	0	18
Annual cost of debt to financial institutions in the short-run (%)	Domestic	4.04	4.55	0	27
	Foreign	4.46	4.40	0	18
Number of markets	Domestic	1.95	1.19	1	5
	Foreign	2.23	1.24	1	5
Proportion of public capital (%)	Domestic	0.99	8.83	0	100
	Foreign	0.63	5.36	0	67

Data Source: ESSE

TABLE B.6. Worker Selection Service

Service is:	Domestic		Foreign	
	Obs. (%)	Mean Pte	Obs. (%)	Mean Pte
Performed by the firm	71.68	23.86	54.06	15.15
Subcontracted	12.02	18.11	42.28	11.96
Not used	16.31	19.58	3.65	9.41
Total	14,105		1,970	

Data Source: ESSE

TABLE B.7. Worker Formation Service

Service is:	Domestic		Foreign	
	Obs. (%)	Mean Pte	Obs. (%)	Mean Pte
Performed by the firm	56.98	24.86	34.16	14.74
Subcontracted	24.52	18.41	61.93	12.73
Not used	18.50	20.50	3.91	17.36
Total	14,105		1,970	

Data Source: ESSE

TABLE B.8. Legal Form

Legal Form	Domestic		Foreign	
	Obs. (%)	Mean Pte	Obs. (%)	Mean Pte
1: Corporations	53.71	18.04	92.34	13.71
2: Limited Liability Companies	38.30	28.33	6.80	12.48
3: Others	7.99	24.22	0.86	9.50
Total	14,105		1,970	

Data Source: ESSE

TABLE B.9. Firm Region

Region	Domestic		Foreign	
	Obs. (%)	Mean Pte	Obs. (%)	Mean Pte
1: Andalucia	8.03	37.13	3.15	9.47
2: Aragon	3.65	24.53	7.77	15.86
3: Asturias	1.94	26.02	1.27	5.41
4: Balearic Islands	1.88	20.31	0.2	13.44
5: Canary Islands	1.58	18.18	1.12	7.14
6: Cantabria	1.13	11.62	2.84	13.38
7: Castilla-La Mancha	5.53	28.80	1.68	9.80
8: Castilla-Leon	4.49	20.83	6.29	11.27
9: Catalonia	20.29	16.91	31.57	12.81
10: Valencian Community	17.28	26.67	6.24	15.37
11: Extremadura	1.03	25.52	0.91	61.93
12: Galicia	5.29	29.56	2.49	18.59
13: Madrid	16.36	15.70	22.64	12.97
14: Murcia	2.74	34.21	0.61	56.42
15: Navarra	1.66	16.40	3.76	17.75
16: Basque Country	5.76	15.10	7.16	8.08
17: La Rioja	1.35	18.43	0.3	25.82
Total	14,105		1,970	

Data Source: ESSE

TABLE B.10. Type of Good

Type	Domestic		Foreign	
	Obs. (%)	Mean Pte	Obs. (%)	Mean Pte
1: Consumer good	23.17	22.94	11.12	12.20
2: Intermediate good	50.34	21.57	67.87	14.01
3: Undefined	26.49	23.79	21.02	13.00
Total	14,105		1,970	

Data Source: ESSE

TABLE B.11. Market Dynamism

Evolution	Domestic		Foreign	
	Obs. (%)	Mean Pte	Obs. (%)	Mean Pte
Expansion	26.13	25.76	34.87	16.17
Stable	52.41	22.24	45.33	12.89
Slump	21.46	19.03	19.80	10.66
Total	14,105		1,970	

Data Source: ESSE

TABLE B.12. Demand Changes

Change in Demand	Domestic		Foreign	
	Obs. (%)	Mean Pte	Obs. (%)	Mean Pte
Demand increase	3.41	23.36	5.13	11.43
Demand decrease	4.75	17.97	6.00	11.05
No change	91.84	22.67	89.00	13.89
Total	14,105		1,970	

Data Source: ESSE

Bibliography

Andersen, E. B. (1970): "Asymptotic Properties of Conditional Maximum-Likelihood Estimators," *Journal of the Royal Statistical Society. Series B (Methodological)*, 32(2), 283–301.

Bentolila, S., and J. J. Dolado (1994): "Labor Flexibility and Wages: Lessons from Spain," *Economic Policy*, 9(18), 53–99.

Blanchard, O., and A. Landier (2002): "The Perverse Effects of Partial Labour Market Reform: Fixed-Term Contracts in France," *The Economic Journal*, 112(480), F214–F244.

Conyon, M. J., S. Girma, S. Thompson, and P. W. Wright (2002): "The Productivity and Wage Effects of Foreign Acquisition in the United Kingdom," *The Journal of Industrial Economics*, 50(1), 85–102.

Dolado, J., C. García-Serrano, and J. Jimeno (2002): "Drawing Lessons from the Boom of Temporary Jobs in Spain," *The Economic Journal*, 112(480), 270–295.

Feliciano, Z., and R. E. Lipsey (1999): "Foreign Ownership and Wages in the United States, 1987 - 1992," NBER Working Papers 6923, National Bureau of Economic Research.

García, A., J. Jaumandreu, and C. Rodríguez (2002): "Innovation and jobs: evidence from manufacturing firms," Discussion paper, Universidad Carlos III de Madrid.

Garibaldi, P., and P. Mauro (2002): "Employment growth. Accounting for the facts," *Economic Policy*, 17(34), 67–113.

Görg, H., E. Strobl, and F. Walsh (2007): "Why Do Foreign-Owned Firms Pay More? The Role of On-the-Job Training," *Review of World Economics*, 143(3), 464–482.

Griffith, R., and H. Simpson (2003): "Characteristics of Foreign-Owned Firms in British Manufacturing," NBER Working Papers 9573, National Bureau of Economic Research, Inc.

Güell, M. (2000): "Fixed-term Contracts and Unemployment: an Efficiency Wage Analysis," Papers 433, Princeton, Department of Economics, Industrial Relations Section.

Hamermesh, D. S., and G. A. Pfann (1996): "Adjustment Costs in Factor Demand," *Journal of Economic Literature*, 34(3), 1264–1292.

Heckman, J. J. (1979): "Sample Selection Bias as a Specification Error," *Econometrica*, 47(1), 153–161.

Honoré, B. E., and E. Kyriazidou (2000): "Panel Data Discrete Choice Models with Lagged Dependent Variables," *Econometrica*, 68(4), 839–874.

Katz, L. F. (1986): "Efficiency Wage Theories: A Partial Evaluation," *NBER Macroeconomics Annual*, 1, 235–276.

Kennan, J. (1979): "The Estimation of Partial Adjustment Models with Rational Expectations," *Econometrica*, 47(6), 1441–1455.

Lindbeck, A., and D. Snower (1988): *The insider-outsider theory of employment and unemployment*. MIT Press, Cambridge, Mass.

Nickell, S., J. Vainiomaki, and S. Wadhwani (1994): "Wages and Product Market Power," *Economica*, 61(244), 457–473.

Nickell, S. J. (1978): "Fixed Costs, Employment and Labour Demand over the Cycle," *Economica*, 45(180), 329–345.

Oswald, A. J. (1985): "The Economic Theory of Trade Unions: An Introductory Survey," *Scandinavian Journal of Economics*, 87(2), 160–93.

Sánchez, R., and L. Toharia (2000): "Temporary Workers and Productivity: The Case of Spain," *Applied Economics*, 32(5), 583–91.

Sargent, T. J. (1978): "Estimation of Dynamic Labor Demand Schedules under Rational Expectations," *The Journal of Political Economy*, 86(6), 1009–1044.

Tobin, J. (1958): "Estimation of Relationships for Limited Dependent Variables," *Econometrica*, 26(1), 24–36.

TABLE 1. Results of the Baseline Estimations

Explanatory Variables	Tobit Model	Heckman Model			
		No Fixed Effects		Fixed Effects	
		Entry Equation	Level-of-use Equation	Entry Equation	Level-of-use Equation
Foreign capital of small firms	-0.0880***	-0.2339***	-0.0446	-0.2339***	-0.0152
	[0.024]	[0.089]	[0.039]	[0.089]	[0.051]
Foreign capital of medium firms	-0.0724***	-0.1050***	-0.0871***	-0.1050***	-0.0583**
	[0.015]	[0.035]	[0.022]	[0.035]	[0.029]
Foreign capital of large firms	-0.0278***	-0.0191	-0.0443***	-0.0191	-0.0372**
	[0.007]	[0.017]	[0.011]	[0.017]	[0.016]
Proportion of foreign capital	-0.0729***	-0.1680***	-0.0522*	-0.1680***	-0.0282
	[0.018]	[0.063]	[0.027]	[0.063]	[0.035]
Small firm	1.0235*	-1.5586	-1.3108	-1.5586	-6.3960***
	[0.608]	[1.883]	[1.062]	[1.883]	[1.464]
Medium firm	2.1636***	2.1077	0.4442	2.1077	-3.9520***
	[0.405]	[1.461]	[0.698]	[1.461]	[0.951]
Prop. of eng. and bachelors	-0.0687***	0.0087	-0.1601***	0.0087	0.0722
	[0.021]	[0.061]	[0.031]	[0.061]	[0.048]
Capital-to-labor ratio	-0.0000***	-0.0000***	-0.0000***	-0.0000***	-0.0000
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
Age	-0.2608***	-0.3267***	-0.3789***	-0.3267***	-0.8515**
	[0.011]	[0.025]	[0.016]	[0.025]	[0.356]
Market dynamism: expansion	1.8751***	3.6933***	1.9467***	3.6933***	1.1412***
	[0.276]	[0.749]	[0.419]	[0.749]	[0.317]
Market dynamism: slump	-1.8145***	-3.3074***	-2.4509***	-3.3074***	-2.4174***
	[0.305]	[0.868]	[0.504]	[0.868]	[0.398]
Demand increase	-0.5245	0.1824	-1.8094**	0.1824	-1.8489***
	[0.562]	[1.762]	[0.912]	[1.762]	[0.631]
Demand decrease	-1.0304**	1.2565	-3.1479***	1.2565	-0.4468
	[0.479]	[1.446]	[0.760]	[1.446]	[0.575]
Public control	-6.4124***	-10.0569***	-11.7366***	-10.0569***	
	[0.607]	[3.328]	[1.280]	[3.328]	
Export propensity	0.0135**	-0.0121	0.0272***	-0.0121	0.0050
	[0.006]	[0.016]	[0.009]	[0.016]	[0.013]
Production of good and services	1.9909***	8.2223***	-0.2457	8.2223***	4.8896***
	[0.192]	[0.492]	[0.303]	[0.492]	[0.652]
Average productivity of labor	-0.0000***	-0.0000***	-0.0000	-0.0000***	-0.0000***
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
Market share	-0.0253***	-0.0566***	-0.0171*	-0.0566***	0.0248*
	[0.007]	[0.020]	[0.010]	[0.020]	[0.014]
Market globalization index	0.0041	0.0354***	-0.0046	0.0354***	-0.0039
	[0.003]	[0.010]	[0.005]	[0.010]	[0.005]
Market atomization index	0.0055*	-0.0161**	0.0235***	-0.0161**	-0.0067
	[0.003]	[0.008]	[0.005]	[0.008]	[0.005]

TABLE 1. Results of the Baseline Estimations (continuation)

Explanatory Variables	Tobit Model	Heckman Model			
		No Fixed Effects		Fixed Effects	
		Entry Equation	Level-of-use Equation	Entry Equation	Level-of-use Equation
Worker formation not used	-1.0096**	-3.0514**	-0.2845	-3.0514**	-0.8424
	[0.443]	[1.217]	[0.685]	[1.217]	[0.672]
Worker formation by the firm	0.3509	-0.9964	1.1029***	-0.9964	0.2165
	[0.285]	[0.870]	[0.415]	[0.870]	[0.429]
Worker selection not used	-3.2199***	-8.7062***		-8.7062***	
	[0.488]	[1.463]		[1.463]	
Worker selection by the firm	-0.7095**	-1.5471		-1.5471	
	[0.322]	[1.031]		[1.031]	
Family business	-0.2020***	-0.2990***		-0.2990***	
	[0.027]	[0.059]		[0.059]	
Cost of debt in the long-run	0.1238***	0.3407***	0.1175**	0.3407***	0.0932**
	[0.033]	[0.095]	[0.052]	[0.095]	[0.041]
Cost of debt in the short-run	0.0038	0.0264	0.0094	0.0264	0.0334
	[0.028]	[0.075]	[0.043]	[0.075]	[0.037]
Labor reform of 1994		1.7912*		1.7912*	
		[0.949]		[0.949]	
Labor reform of 1997		-3.1850***		-3.1850***	
		[0.944]		[0.944]	
Labor reform of 2001		-9.1093***		-9.1093***	
		[1.004]		[1.004]	
Lambda			-6.3686**		-0.6443
			[2.615]		[2.491]

Number of observations = 16,075. All regressions include industry, region and time dummies and variables of legal type, type of good and number of markets. Tobit and entry equations also include dummies for worker selection and outsourcing. Standard errors robust to heteroskedasticity in brackets.
* significant at 10%; ** significant at 5%; *** significant at 1%.

TABLE 2. Average effect of a change of pfk to 50%

Explanatory Variables	Tobit Model	Heckman Model			
		No Fixed Effects		Fixed Effects	
		Entry Equation	Level-of-use Equation	Entry Equation	Level-of-use Equation
Domestic to foreign					
All firms	-2.8970***	-8.5241***	-1.6361	-8.5241***	-0.1310
	[0.618]	[2.424]	[1.117]	[2.424]	[1.487]
Small firms	-3.1097***	-10.7621***	-0.7769	-10.7621***	1.0780
	[0.764]	[2.948]	[1.486]	[2.948]	[2.019]
Medium firms	-3.1521***	-5.2836***	-3.9008***	-5.2836***	-2.3778*
	[0.545]	[1.964]	[0.909]	[1.964]	[1.229]
Large firms	-1.6592***	-1.3455	-2.6998***	-1.3455	-2.5104**
	[0.466]	[1.343]	[0.833]	[1.343]	[1.065]
Foreign to domestic					
All firms	0.0962	1.0053	0.0007	1.0053	-1.1740
	[0.425]	[1.469]	[0.765]	[1.469]	[0.915]
Small firms	0.4980	4.9910**	-1.5663	4.9910**	-3.2587**
	[0.460]	[2.342]	[1.161]	[2.342]	[1.478]
Medium firms	0.6968	2.6214	0.7506	2.6214	-1.0798
	[0.439]	[1.892]	[0.815]	[1.892]	[0.995]
Large firms	-0.2659	-0.4534	-0.1559	-0.4534	-0.9995
	[0.463]	[1.298]	[0.799]	[1.298]	[0.992]

* significant at 10%; ** significant at 5%; *** significant at 1%

TABLE 3. Results of Specification II

Explanatory Variables	Tobit Model	Heckman Model			
		No Fixed Effects		Fixed Effects	
		Entry Equation	Level-of-use Equation	Entry Equation	Level-of-use Equation
Foreign capital of small firms	-0.0599	-0.1815	-0.0963**	-0.1815	-0.0631
	[0.049]	[0.130]	[0.045]	[0.130]	[0.055]
Foreign capital of medium firms	-0.0780***	-0.0931**	-0.1014***	-0.0931**	-0.0585*
	[0.016]	[0.037]	[0.023]	[0.037]	[0.030]
Foreign capital of large firms	-0.0409***	-0.0115	-0.0581***	-0.0115	-0.0186
	[0.007]	[0.017]	[0.012]	[0.017]	[0.016]
Proportion of foreign capital	-0.0590*	-0.1316	-0.0881***	-0.1316	-0.0516
	[0.033]	[0.088]	[0.031]	[0.088]	[0.037]
Small firm	0.9958	-0.7802	-1.4251	-0.7802	-5.9498***
	[0.631]	[1.822]	[1.050]	[1.822]	[1.446]
Medium firm	2.0306***	2.5124*	0.3737	2.5124*	-3.7230***
	[0.418]	[1.465]	[0.688]	[1.465]	[0.941]
Prop. of eng. and bachelors	-0.0761***	-0.0033	-0.1580***	-0.0033	0.0706
	[0.022]	[0.062]	[0.033]	[0.062]	[0.048]
Capital-to-labor ratio	-0.0000***	-0.0000***	-0.0000***	-0.0000***	-0.0000
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
Age	-0.2728***	-0.3282***	-0.4081***	-0.3282***	-0.8872**
	[0.011]	[0.025]	[0.016]	[0.025]	[0.354]
Market dynamism: expansion	1.6623***	3.8574***	1.8334***	3.8574***	1.1416***
	[0.270]	[0.791]	[0.424]	[0.791]	[0.318]
Market dynamism: slump	-1.8962***	-3.1302***	-2.2190***	-3.1302***	-2.3136***
	[0.322]	[0.811]	[0.509]	[0.811]	[0.395]
Demand increase	-0.5343	0.2395	-1.8287*	0.2395	-1.6762***
	[0.604]	[1.777]	[0.960]	[1.777]	[0.650]
Demand decrease	-0.9728*	1.1625	-3.2605***	1.1625	-0.3092
	[0.505]	[1.487]	[0.760]	[1.487]	[0.577]
Public control	-7.2395***	-8.9891***	-9.0600***	-8.9891***	
	[0.764]	[2.806]	[1.158]	[2.806]	
Export propensity	0.0170***	-0.0044	0.0441***	-0.0044	0.0029
	[0.006]	[0.017]	[0.010]	[0.017]	[0.013]
Production of good and services	2.1104***	8.5320***	-0.1243	8.5320***	5.2686***
	[0.192]	[0.496]	[0.301]	[0.496]	[0.650]
Average productivity of labor	-0.0000***	-0.0001***	0.0000	-0.0001***	-0.0000***
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
Market share	-0.0239***	-0.0488**	-0.0161	-0.0488**	0.0232*
	[0.007]	[0.020]	[0.011]	[0.020]	[0.013]
Market globalization index	0.0008	0.0249**	-0.0052	0.0249**	-0.0052
	[0.003]	[0.011]	[0.005]	[0.011]	[0.006]
Market atomization index	0.0071**	-0.0182**	0.0258***	-0.0182**	-0.0066
	[0.003]	[0.008]	[0.005]	[0.008]	[0.005]

TABLE 3. Results of Specification II (continuation)

Explanatory Variables	Tobit Model	Heckman Model			
		No Fixed Effects		Fixed Effects	
		Entry Equation	Level-of-use Equation	Entry Equation	Level-of-use Equation
Worker formation not used	-0.5917	-1.7305	0.3622	-1.7305	-0.7627
	[0.473]	[1.262]	[0.746]	[1.262]	[0.705]
Worker formation by the firm	0.3737	-1.0485	1.3947***	-1.0485	0.1675
	[0.296]	[0.894]	[0.432]	[0.894]	[0.436]
Worker selection not used	-3.2973***	-8.7972***		-8.7972***	
	[0.531]	[1.469]		[1.469]	
Worker selection by the firm	-0.5973*	-2.0616*		-2.0616*	
	[0.344]	[1.148]		[1.148]	
Family business	-0.0736	-0.2295		-0.2295	
	[0.096]	[0.169]		[0.169]	
Cost of debt in the long-run	0.1227***	0.3433***	0.1041**	0.3433***	0.0867**
	[0.032]	[0.096]	[0.050]	[0.096]	[0.039]
Cost of debt in the short-run	0.0309	0.0307	0.0567	0.0307	0.0377
	[0.028]	[0.075]	[0.043]	[0.075]	[0.036]
Labor reform of 1994		2.1014**		2.1014**	
		[1.068]		[1.068]	
Labor reform of 1997		-3.2123***		-3.2123***	
		[0.998]		[0.998]	
Labor reform of 2001		-8.6825***		-8.6825***	
		[1.008]		[1.008]	
Lambda			-6.9912***		-0.1226
			[2.116]		[2.430]

Number of observations = 16,075. All regressions include industry, region and time dummies and variables of legal type, type of good and number of markets. Tobit and entry equations also include dummies for worker selection and outsourcing. Standard errors robust to heteroskedasticity in brackets.
* significant at 10%; ** significant at 5%; *** significant at 1%.

TABLE 4. Average effect of a change of pfk to 50%. Specification II.

Explanatory Variables	Tobit Model	Heckman Model			
		No Fixed Effects		Fixed Effects	
		Entry Equation	Level-of-use Equation	Entry Equation	Level-of-use Equation
Domestic to foreign					
All firms	-1.8250	-6.9554	-3.7162***	-6.9554	-1.8511
	[1.944]	[4.597]	[1.368]	[4.597]	[1.696]
Small firms	-1.4244	-8.7812	-3.5503*	-8.7812	-1.6894
	[2.672]	[6.226]	[1.861]	[6.226]	[2.310]
Medium firms	-3.3706***	-4.7293**	-4.8574***	-4.7293**	-2.7650**
	[0.642]	[2.047]	[0.960]	[2.047]	[1.315]
Large firms	-2.1550***	-0.6875	-3.2602***	-0.6875	-1.5930
	[0.471]	[1.331]	[0.844]	[1.331]	[1.055]
Foreign to domestic					
All firms	0.4760	1.4109	0.6034	1.4109	-1.6324*
	[0.428]	[1.485]	[0.763]	[1.485]	[0.918]
Small firms	0.4581	5.1310**	-1.5205	5.1310**	-3.4812**
	[0.472]	[2.394]	[1.117]	[2.394]	[1.516]
Medium firms	0.7334*	2.9535	0.9105	2.9535	-1.3340
	[0.437]	[1.910]	[0.820]	[1.910]	[1.033]
Large firms	0.3530	0.0325	0.6917	0.0325	-1.5686
	[0.465]	[1.305]	[0.803]	[1.305]	[0.983]

* significant at 10%; ** significant at 5%; *** significant at 1%

TABLE 5. Specification II: Effect of pfk for different groups of firms.

Explanatory Variables	Tobit Model	Heckman Model			
		No Fixed Effects		Fixed Effects	
		Entry Equation	Level-of-use Equation	Entry Equation	Level-of-use Equation
Capital-to-labor ratio					
First quartile	-0.0727	-0.1647	-0.1270***	-0.1647	-0.0892
	[0.054]	[0.132]	[0.045]	[0.132]	[0.054]
Second quartile	-0.0654	-0.1457	-0.1019***	-0.1457	-0.0634
	[0.041]	[0.107]	[0.038]	[0.107]	[0.045]
Third quartile	-0.0585**	-0.1241	-0.0790***	-0.1241	-0.0413
	[0.027]	[0.076]	[0.027]	[0.076]	[0.033]
Fourth quartile	-0.0393***	-0.0920**	-0.0466***	-0.0920**	-0.0144
	[0.012]	[0.040]	[0.016]	[0.040]	[0.020]
Age					
Less than 10 years old	-0.0972**	-0.1535	-0.1326***	-0.1535	-0.0750
	[0.049]	[0.100]	[0.039]	[0.100]	[0.048]
Between 11 and 20 years old	-0.0611	-0.1496	-0.1006***	-0.1496	-0.0572
	[0.040]	[0.109]	[0.036]	[0.109]	[0.043]
Between 21 and 30 years old	-0.0499**	-0.1289	-0.0801***	-0.1289	-0.0464
	[0.025]	[0.082]	[0.027]	[0.082]	[0.033]
More than 31 years old	-0.0147	-0.0872	-0.0202	-0.0872	-0.0178
	[0.013]	[0.058]	[0.020]	[0.058]	[0.025]
Market dynamism: expansion	-0.0570*	-0.1416	-0.0791***	-0.1416	-0.0394
	[0.035]	[0.094]	[0.028]	[0.094]	[0.034]
Market dynamism: slump	-0.0449	-0.1273	-0.0984***	-0.1273	-0.0582
	[0.034]	[0.108]	[0.034]	[0.108]	[0.039]
Demand increase	-0.0802**	-0.0995	-0.1106***	-0.0995	-0.0638*
	[0.032]	[0.080]	[0.033]	[0.080]	[0.034]
Demand decrease	-0.0499	-0.0772	-0.0889***	-0.0772	-0.0771**
	[0.030]	[0.092]	[0.032]	[0.092]	[0.036]
Average productivity of labor					
First quartile	-0.063	-0.1715	-0.1292***	-0.1715	-0.1015*
	[0.051]	[0.136]	[0.044]	[0.136]	[0.054]
Second quartile	-0.0680*	-0.1462	-0.0999***	-0.1462	-0.0743*
	[0.038]	[0.100]	[0.036]	[0.100]	[0.044]
Third quartile	-0.0645**	-0.1260*	-0.0788***	-0.1260*	-0.0457
	[0.027]	[0.069]	[0.027]	[0.069]	[0.033]
Fourth quartile	-0.0408**	-0.0828*	-0.0468**	-0.0828*	0.0118
	[0.018]	[0.049]	[0.019]	[0.049]	[0.023]

TABLE 5. Specification II: Effect of pfk for different groups of firms (continuation).

Explanatory Variables	Tobit Model	Heckman Model			
		No Fixed Effects		Fixed Effects	
		Entry Equation	Level-of-use Equation	Entry Equation	Level-of-use Equation
Market globalization index					
Equal to zero	-0.0645*	-0.1565	-0.0968***	-0.1565	-0.0600
	[0.038]	[0.099]	[0.035]	[0.099]	[0.042]
Between 0 and 100	-0.0440**	-0.0606	-0.0617***	-0.0606	-0.0276
	[0.019]	[0.047]	[0.020]	[0.047]	[0.025]
Equal to 100	-0.0356**	-0.0278	-0.0563***	-0.0278	-0.0206
	[0.017]	[0.050]	[0.019]	[0.050]	[0.023]
Market atomization index					
Equal to zero	-0.0572*	-0.1094	-0.0856***	-0.1094	-0.0472
	[0.030]	[0.081]	[0.029]	[0.081]	[0.035]
Between 0 and 100	-0.0549*	-0.1259	-0.0764**	-0.1259	-0.0466
	[0.032]	[0.081]	[0.030]	[0.081]	[0.037]
Equal to 100	-0.0673	-0.2199*	-0.1033**	-0.2199*	-0.0723
	[0.046]	[0.121]	[0.044]	[0.121]	[0.051]
Worker formation not used	-0.0796**	-0.1880*	-0.0486	-0.1880*	-0.0538
	[0.040]	[0.097]	[0.053]	[0.097]	[0.056]
Worker formation by the firm	-0.0533***	-0.0834*	-0.1080***	-0.0834*	-0.0664
	[0.015]	[0.043]	[0.036]	[0.043]	[0.043]
Worker selection not used	-0.0663*	-0.1479*	-0.0763	-0.1479*	-0.0618
	[0.036]	[0.089]	[0.049]	[0.089]	[0.054]
Worker selection by the firm	-0.0519***	-0.0722**	-0.0981***	-0.0722**	-0.0588
	[0.013]	[0.030]	[0.033]	[0.030]	[0.040]
Family business					
Equal to zero	-0.0815***	-0.1186**	-0.0754***	-0.1186**	-0.0350
	[0.015]	[0.049]	[0.021]	[0.049]	[0.028]
Greater than zero	-0.034	-0.1463	-0.1033**	-0.1463	-0.0715
	[0.066]	[0.158]	[0.044]	[0.158]	[0.050]

* significant at 10%; ** significant at 5%; *** significant at 1%.

TABLE 6. Results of Specification III

Explanatory Variables	Tobit Model	Heckman Model			
		No Fixed Effects		Fixed Effects	
		Entry Equation	Level-of-use Equation	Entry Equation	Level-of-use Equation
Foreign capital of small firms	-0.0791***	-0.2559***	-0.0191	-0.2559***	0.0041
	[0.026]	[0.093]	[0.042]	[0.093]	[0.050]
Foreign capital of medium firms	-0.0644***	-0.1115***	-0.0666***	-0.1115***	-0.0479*
	[0.016]	[0.037]	[0.024]	[0.037]	[0.029]
Foreign capital of large firms	-0.0255***	-0.0226	-0.0357***	-0.0226	-0.0300*
	[0.007]	[0.014]	[0.011]	[0.014]	[0.016]
Foreign capital	-0.0662***	-0.1869***	-0.0312	-0.1869***	-0.0128
	[0.020]	[0.067]	[0.030]	[0.067]	[0.035]
Lambda			-5.3719**		-0.1013
			[2.646]		[2.497]

Number of observations = 15,635. All regressions include industry, region and time dummies, and other covariates included in the baseline specification. Standard errors robust to heteroskedasticity in brackets. * significant at 10%; ** significant at 5%; *** significant at 1%

TABLE 7. Average effect of a change of pfk to 50%. Specification III.

Explanatory Variables	Tobit Model	Heckman Model			
		No Fixed Effects		Fixed Effects	
		Entry Equation	Level-of-use Equation	Entry Equation	Level-of-use Equation
Domestic to foreign					
All firms	-2.7258***	-9.2142***	-0.9681	-9.2142***	0.5062
	[0.664]	[2.570]	[1.228]	[2.570]	[1.478]
Small firms	-2.9232***	-11.3493***	-0.1959	-11.3493***	1.7410
	[0.798]	[3.067]	[1.595]	[3.067]	[1.975]
Medium firms	-2.8814***	-5.6933***	-3.1067***	-5.6933***	-1.9804
	[0.595]	[2.108]	[0.977]	[2.108]	[1.246]
Large firms	-1.5433***	-1.8072	-2.0582**	-1.8072	-2.1451**
	[0.521]	[1.382]	[0.880]	[1.382]	[1.092]
Foreign to domestic					
All firms	0.2853	0.7289	0.6322	0.7289	-1.0124
	[0.463]	[1.577]	[0.834]	[1.577]	[1.011]
Small firms	0.6597	4.6304*	-1.0523	4.6304*	-3.4013**
	[0.494]	[2.502]	[1.238]	[2.502]	[1.541]
Medium firms	0.8290*	2.1183	1.2714	2.1183	-0.9439
	[0.480]	[2.050]	[0.863]	[2.050]	[1.127]
Large firms	-0.0384	-0.5841	0.5389	-0.5841	-0.7971
	[0.500]	[1.378]	[0.879]	[1.378]	[1.071]

* significant at 10%; ** significant at 5%; *** significant at 1%

TABLE 8. Results of Specification IV

Explanatory Variables	Tobit Model	Heckman Model			
		No Fixed Effects		Fixed Effects	
		Entry Equation	Level-of-use Equation	Entry Equation	Level-of-use Equation
Foreign capital of small firms	-0.0778***	-0.2248**	-0.0315	-0.2248**	0.0297
	[0.026]	[0.105]	[0.044]	[0.105]	[0.058]
Foreign capital of medium firms	-0.0706***	-0.0962**	-0.0919***	-0.0962**	-0.0220
	[0.016]	[0.041]	[0.024]	[0.041]	[0.033]
Foreign capital of large firms	-0.0261***	-0.0106	-0.0462***	-0.0106	-0.0302*
	[0.008]	[0.018]	[0.013]	[0.018]	[0.018]
Foreign capital	-0.0663***	-0.1599**	-0.0463	-0.1599**	0.0059
	[0.020]	[0.075]	[0.030]	[0.075]	[0.039]
Dummy of demand increase (lag)	0.5784	2.6352	1.2842***	2.6352	0.4396
	[0.596]	[1.833]	[0.448]	[1.833]	[0.659]
Dummy of demand decrease (lag)	-1.5961***	-1.0429	-2.6665***	-1.0429	-0.4382
	[0.508]	[1.696]	[0.545]	[1.696]	[0.608]
Prod. of good and services (lag)	1.6399***	7.2483***	-0.4081	7.2483***	1.5751**
	[0.202]	[0.547]	[0.340]	[0.547]	[0.633]
Lambda			-3.080		-1.5197
			[2.577]		[2.418]

Number of observations = 12,900. All regressions include industry, region and time dummies, and other covariates included in the baseline specification. Standard errors robust to heteroskedasticity in brackets. * significant at 10%; ** significant at 5%; *** significant at 1%

TABLE 9. Average effect of a change of pfk to 50%. Specification IV.

Explanatory Variables	Tobit Model	Heckman Model			
		No Fixed Effects		Fixed Effects	
		Entry Equation	Level-of-use Equation	Entry Equation	Level-of-use Equation
Domestic to foreign					
All firms	-2.5964	-8.3540	-1.1051	-8.3540	1.4019
	[0.677]	[2.797]	[1.275]	[2.797]	[1.712]
Small firms	-2.7000***	-10.7097***	0.1562	-10.7097***	2.8744
	[0.836]	[3.411]	[1.753]	[3.411]	[2.410]
Medium firms	-3.0809***	-4.7725**	-4.0306***	-4.7725**	-0.7108
	[0.600]	[2.225]	[0.972]	[2.225]	[1.353]
Large firms	-1.5843***	-0.7874	-2.9184***	-0.7874	-2.0137
	[0.526]	[1.445]	[0.910]	[1.445]	[1.231]
Foreign to domestic					
All firms	0.0647***	1.2625***	-0.3334	1.2625***	-1.1865
	[0.471]	[1.680]	[0.849]	[1.680]	[1.003]
Small firms	0.3533	5.6481**	-2.4801*	5.6481**	-3.8331**
	[0.512]	[2.671]	[1.375]	[2.671]	[1.879]
Medium firms	0.6828	2.9984	0.4470	2.9984	-1.7286
	[0.482]	[2.220]	[0.870]	[2.220]	[1.063]
Large firms	-0.3071	-0.3853	-0.4650	-0.3853	-0.6848
	[0.516]	[1.436]	[0.890]	[1.436]	[1.068]

* significant at 10%; ** significant at 5%; *** significant at 1%

TABLE 10. Results of Specification V

Explanatory Variables	Tobit Model	Heckman Model	
		Entry Equation	Level-of-use Equation
Foreign capital of small firms	-4.0731***	-15.8442***	0.2147
	[1.071]	[3.578]	[1.906]
Foreign capital of medium firms	-4.2181***	-7.8705***	-5.4656***
	[0.530]	[1.870]	[1.040]
Foreign capital of large firms	-1.6442***	-0.8203	-2.9544***
	[0.370]	[1.132]	[0.784]
Proportion of foreign capital	-3.5960***	-11.4113***	-1.3882
	[0.688]	[2.305]	[1.233]
Small firm	1.1409*	-1.8078	-1.1042
	[0.607]	[1.864]	[0.982]
Medium firm	2.2635***	1.8875	0.8032
	[0.403]	[1.442]	[0.698]
Changes nationality	1.9956***	1.1427	3.0817***
	[0.450]	[1.494]	[0.749]
Prop. of eng. and bachelors	-0.0738***	0.0070	-0.1658***
	[0.021]	[0.061]	[0.035]
Capital-to-labor ratio	-0.0000***	-0.0000***	-0.0000***
	[0.000]	[0.000]	[0.000]
Age	-0.2603***	-0.3405***	-0.3781***
	[0.011]	[0.026]	[0.014]
Market dynamism: expansion	1.8736***	3.6144***	1.9471***
	[0.276]	[0.749]	[0.417]
Market dynamism: slump	-1.8028***	-3.2813***	-2.4520***
	[0.305]	[0.867]	[0.489]
Demand increase	-0.6248	0.2070	-1.9323**
	[0.561]	[1.762]	[0.917]
Demand decrease	-1.0399**	1.2777	-3.1809***
	[0.480]	[1.446]	[0.825]
Public control	-6.3696***	-9.7262***	-11.7342***
	[0.610]	[3.310]	[1.540]
Export propensity	0.0122**	-0.0174	0.0255***
	[0.006]	[0.016]	[0.009]
Production of good and services	1.9506***	7.9972***	-0.2280
	[0.192]	[0.494]	[0.282]
Average productivity of labor	-0.0000***	-0.0000***	-0.0000
	[0.000]	[0.000]	[0.000]
Market share	-0.0265***	-0.0588***	-0.0192*
	[0.007]	[0.020]	[0.011]
Market globalization index	0.0039	0.0355***	-0.0049
	[0.003]	[0.010]	[0.005]
Market atomization index	0.0054*	-0.0168**	0.0230***
	[0.003]	[0.008]	[0.005]

TABLE 10. Results of Specification V (continuation)

Explanatory Variables	Tobit Model	Heckman Model	
		Entry Equation	Level-of-use Equation
Worker formation not used	-3.1520***	-8.8987***	
	[0.487]	[1.460]	
Worker formation by the firm	-0.6554**	-1.5714	
	[0.321]	[1.025]	
Worker selection not used	-1.0055**	-3.0267**	-0.4344
	[0.443]	[1.215]	[0.634]
Worker selection by the firm	0.3496	-1.1023	1.1190**
	[0.285]	[0.870]	[0.447]
Family business	-0.2025***	-0.3273***	
	[0.027]	[0.060]	
Cost of debt in the long-run	0.1241***	0.3385***	0.1172**
	[0.033]	[0.095]	[0.049]
Cost of debt in the short-run	0.0061	0.0304	0.0125
	[0.028]	[0.075]	[0.041]
Labor reform of 1994		1.7273*	
		[0.949]	
Labor reform of 1997		-3.2866***	
		[0.944]	
Labor reform of 2001		-9.0959***	
		[1.003]	
Lambda			-3.5821***
			[1.136]

Number of observations = 16,075. All regressions include industry, region and time dummies and variables of legal type, type of good and number of markets. Tobit and entry equations also include dummies for worker selection and outsourcing.
Standard errors robust to heteroskedasticity in brackets
* significant at 10%; ** significant at 5%; *** significant at 1%

CHAPTER 2

Correcting the bias in the estimation of a dynamic ordered probit with fixed effects of self-assessed health status[1]

ABSTRACT. This paper considers the estimation of a dynamic ordered probit with fixed effects, with an application to self-assessed health status. The well-known estimation problem of this kind of models when T is not very large is specially severe in our model because it contains two fixed effects: one in the linear index equation and one in the cut points. These two fixed effects, instead of only one as usually done, are implied by the potential existence of heterogeneity in both unobserved health status and reporting behavior. The contributions of this paper are twofold. Firstly this paper contributes to the recent literature on bias correction in nonlinear panel data models by applying and studying the finite sample properties of two of the existing proposals to the ordered probit case. The most direct and easily applicable correction to our model is not the best one and still has important biases in our sample sizes. Secondly, we contribute to the literature that studies the determinants of Self-Assessed Health measures by applying the previous analysis on estimation methods to the British Household Panel Survey.

1. Introduction

The estimation of nonlinear panel data models with fixed effects is known to be problematic with the panels usually available, since they do not have a very large number of periods. This is even more severe when estimating dynamic models, like the dynamic ordered probit model. This incidental parameters problem is reflected in the inconsistency of standard estimators like the maximum likelihood estimator (MLE) when the number of individuals N goes to infinity and T is fixed. Even when T goes to infinity, if it does not grow faster than N, the asymptotic normal distribution is not centered at zero due to the bias coming from the incidental parameters. Moreover, this problem results in large finite sample biases of the

[1]This chapter is based on Carro and Traferri (2009).

MLE when using panels where T is not very large. The dynamic ordered probit model is not an exception to this, specially if it contains more than one individual specific parameter, as in our case.

An important part of the research on microeconometrics in recent years has been concerned with finding a solution to this problem, by developing bias-adjusted methods to estimate those models. Given this fast growing literature, there are several bias correction methods we could consider to estimate our model. These methods can be grouped in three approaches.[2] The first one is to construct an analytical or numerical bias correction of a fixed effect estimator. Hahn and Newey (2004), Hahn and Kuersteiner (2004) and Fernandez-Val (2009), for example, take this approach to the problem. The second approach is to correct the bias in moment equations. An example of this is Carro (2007), which uses an estimator of this type to correct the bias in dynamic binary choice models. The third approach is to correct the objective function. Arellano and Hahn (2006) and Bester and Hansen (2009) take this approach, with the latter including an application to a dynamic ordered probit model.

Asymptotically all of the above methods reduce the order of the bias of the MLE from the standard $O(T^{-1})$ to $O(T^{-2})$. Therefore, from this perspective we could use any of the methods developed for dynamic models. A second criteria to choose among the several alternatives is to check the easiness of implementation to our model. From this criteria the estimator that corrects the objective function using a penalty term based on a product of the sample scores and Hessian can be directly applied without modification to our specific model. Bester and Hansen (2009) refer to it as the HS penalty. In contrast with the direct applicability of this estimator, others are computationally more difficult and require some transformation to be applied to our model, specially because our model contains two fixed effects instead of one as usually is the case in binary choice models. This does not mean that other methods cannot be applied nor that we do not know their theoretical properties. They have been developed for a quite general class of nonlinear panel data models with fixed effects.

A third and more important criteria is the finite sample performance of the method when estimating our model with the sample size we have. The incidental

[2]See Arellano and Hahn (2007) for a good review of this literature, detailed references and a general framework in which the various approaches can be included.

parameters problem can be seen as a finite sample bias problem in panel data context. The incidental parameters problem is not very important when T is large relative to N. However, since our panel does not have a very large number of periods it is reasonable to wonder whether the asymptotic properties when T goes to infinity are a good approximation to our finite sample. Given this, we should evaluate the finite sample performance of the available methods we could use to estimate our model. As usual, this comparison is done through Monte Carlo experiments. Bester and Hansen (2009) do not compare the finite sample properties of the method they use with others for the ordered probit case because many of the other methods will require some derivation to get the specific correction for this case. They, however, make such a comparison using a static and a dynamic logit model. Also, Carro (2007) and Fernandez-Val (2009) make Monte Carlo experiments for logit and probit models with different sample sizes. The Monte Carlo experiments made in these three papers allow us to compare a wide range of methods for the dynamic logit and probit models. From all these comparisons we can conclude that the HS penalty approach is clearly not the best one. We can also conclude that for sample sizes with T smaller than 13 the reminding bias when using HS could still be significant, specially for the ordered probit Bester and Hansen (2009) simulate. This result is also confirmed in our simulations. Given this and that our empirical application has $T = 13$, some other of the proposed methods should be considered, in addition to the HS penalty approach. Interesting candidates are the corrections discussed by Fernandez-Val (2009) and Carro (2007) since they are both equally superior to other methods in the relevant existing Monte Carlo experiments. In this paper we derive explicit formulas of the modified MLE used in Carro (2007) for the model considered here, evaluate its finite sample performance and compare it with the HS penalty estimator. This exercise is a main contribution of this paper since, as Arellano and Hahn (2007) point out in their conclusions, more research is needed to know "how well each of the methods recently proposed work for other specific models and data set of interest in applied econometrics." Also, Greene and Hensher (2010) comment on the lack of studies about the applicability to ordered choice models of the recent proposals for bias reduction estimators in binary choice models.

Self-assessed health (SAH) has been used as a proxy for true overall individual health status in many socioeconomic studies. Moreover, it has been shown

to be a good predictor of mortality and of subsequent demand of medical care (see for example van Doorslaer, Koolman and Jones (2004)). Motivated by this importance and the high observed persistence in health outcomes, Contoyannis, Jones, and Rice (2004) study the dynamics and effects of socioeconomic variables on SAH for the British Household Panel Survey. Among other aims, they try to know the relative contribution of state dependence and unobserved heterogeneity in explaining the observed persistence in SAH. Given that SAH is a categorical variable they use a dynamic ordered probit model, and they take a random effects approach to control for unobserved heterogeneity in the level equation.

In addition to accounting for unobserved factors that affect health status (index shift), here we also have to take into account the possible heterogeneity in reporting behavior (cut-point shift). The cut-point shifts occur if individuals use different thresholds when assessing their health and reporting it in the SAH categorical variable, so that they report a different value of SAH even though having the same level of true health.[3] To control for these two unobserved factors, which are possibly correlated with other explanatory variables and between each other, we include individual effects not only in the levels of the ordered probit but also in the cut points. Given that in discrete choice models we can only identify differences up to scale, in addition to the normalization in the errors we have to normalize one of the cut-point shifters (or the index shifter). This means that we cannot separately identify the two sources of heterogeneity. We can, nonetheless, correctly control for these two sources of heterogeneity in the estimation of our model. In contrast, a model that only allows for one individual effect (usually placed in the index equation) and which imposes homogeneity in the other shifters will almost always give incorrect estimates and inference if the two afore mentioned sources of heterogeneity are present and correlated with other explanatory variables.

[3]Lindeboom and Van Doorslaer (2004) present a model where unobserved true health (a continuous variable) determines self-reported health (a categorical and ordered variable), through a series of thresholds. True health, in turn, depends on observable individual characteristics and an index shift. The authors allow cut-points and indexes to differ for different subgroups of the population (they present four groups related with the language or cultural background of the respondent: French only, English only, French and English, and Other), estimate the proposed model using data from the Canadian National Population Health Survey, and reject the hypothesis that the cut-points and index effects are the same for the four subgroups. Therefore, these authors find evidence of the existence of these two different kinds of shifts.

As it happens with one individual effect, we could take a 'random effects' approach. However, this approach has the drawback of imposing either independence, or a specific and restrictive functional form for the relation between the unobserved heterogeneity and other explanatory variables. It also has the drawback of having to deal with the so-called initial conditions problem. Taking a 'fixed effects' approach we leave unrestricted the joint distribution of the two individual effects and their correlation with the explanatory variables. Moreover, there is no initial conditions problem. Despite these advantages, there have been only few applications in health economics of nonlinear panel models with fixed effects, as can be seen by reading Jones's (2007) handbook chapter. This is due to the incidental parameters problem addressed by this paper and the related literature. The estimation of our model and the comparison with random effects estimates show that there is higher state dependence effects and that it matters to flexibly account for more permanent unobserved heterogeneity.

The rest of the paper proceeds as follows. We first present our model and its estimation problems. We comment on the possible solutions from the nonlinear bias correction literature for nonlinear panel data models with fixed effects. We use simulations to evaluate the finite sample performance of two of the alternatives and use this as final criteria for choosing our estimator. In Section 3, we apply all that to the study self-assessed health status in the British Household Panel Survey. There we first present the data and variables we include in our model. The estimates and comments on them follow. Last section concludes.

2. The Model and Estimation Method

We consider a dynamic panel data ordered probit with fixed effects:[4]

$$h_{it}^{*} = \alpha_i + \rho_1 \mathbf{1}\left(h_{i,t-1} = 1\right) + \rho_{-1} \mathbf{1}\left(h_{i,t-1} = -1\right) + x_{it}'\beta + \varepsilon_{it}, \tag{7}$$

for $i = 1, \ldots, N$ and $t = 0, \ldots, T$, where h_{it}^{*} is the latent variable (e.g. health status), and the observed variable (h_{it}) is determined according to the following

[4]In actuality, α_i is equal to the fixed effect plus the coefficient of the missing category ($h_{i,t-1} = 0$). In other words, fixed effects can be identified up to a constant.

thresholds:

$$h_{it} = \begin{cases} -1 & \text{if} \quad h_{it}^* < -c_i \\ 0 & \text{if} \quad -c_i < h_{it}^* \leq 0 \\ 1 & \text{if} \quad h_{it}^* > 0 \end{cases} \tag{8}$$

For instance, in our empirical application, $h_{it} = -1$ corresponds to poor health, $h_{it} = 0$ to fair health and $h_{it} = 1$ to good health. α_i and c_i are the model's fixed effects, and $\varepsilon_{it} \underset{iid}{\sim} N(0,1)$. Note that in addition to the usual scale normalization in discrete choice models, here we are also normalizing one of the two cut points to be zero. The, somehow more conventional, normalization of setting the intercept in the linear index equal to zero is not available to us because with the fixed effects approach the distribution of the intercept, including its mean, is unrestricted. An alternative normalization is to put the two fixed effects in the two cut points and leave the linear index equation without any intercept.

From this discussion on normalization it is clear that it is not possible to separately identify individual effects affecting only h_{it}^* from the individual effects affecting the cut points. Having only the fixed effect in the linear index (α_i) will also allow for heterogeneity in the cut points, but in a very restrictive way. In particular, by introducing only one individual effect (α_i), we would be assuming that the unobserved heterogeneity must have effects of opposite sign in $\Pr(h_{it} = 1)$ and $\Pr(h_{it} = -1)$; and also we would be restricting how these two effects differ in magnitude for all individuals. Having two fixed effects as in (8), we are not imposing any restrictions on the cut-point shifts as well as on the index shift.

From (7), (8) and the assumption about ε_{it}, we have that

$$\begin{aligned} &\Pr(h_{it} = -1 | x_{it}, h_{it-1}, c_i, \alpha_i) = 1 - \Phi(c_i + \mu_{it}) \\ &\Pr(h_{it} = 0 | x_{it}, h_{it-1}, c_i, \alpha_i) = \Phi(c_i + \mu_{it}) - \Phi(\mu_{it}) \\ &\Pr(h_{it} = 1 | x_{it}, h_{it-1}, c_i, \alpha_i) = 1 - \Pr(h_{it} = -1|.) - \Pr(h_{it} = 0|.) = \Phi(\mu_{it}) \end{aligned} \tag{9}$$

where

$$\mu_{it} = \alpha_i + \rho_1 \mathbf{1}(h_{i,t-1} = 1) + \rho_{-1} \mathbf{1}(h_{i,t-1} = -1) + x_{it}'\beta \tag{10}$$

Conditioning on the first observation, the log-likelihood is:

$$l(\rho_1, \rho_{-1}, \beta, \alpha, \mathbf{c}) = \sum_{i=1}^{N} \sum_{t=1}^{T} \{\mathbf{1}\{h_{it} = -1\} \log[1 - \Phi(c_i + \mu_{it})] +$$

$$\mathbf{1}\{h_{it} = 0\} \log[\Phi(c_i + \mu_{it}) - \Phi(\mu_{it})] + \mathbf{1}\{h_{it} = 1\} \log[\Phi(\mu_{it})]\}, \tag{11}$$

2.1. Estimation problem and possible solutions. Using standard MLE to estimate models like (8) is well known to be biased, since we do not have a large number of periods. The MLE is inconsistent when T is not going to infinity because the fixed effects are acting as incidental parameters. Furthermore, existing Monte Carlo experiments with nonlinear models similar to this shows that the MLE has large bias. In fact, simulations of a dynamic ordered probit in Bester and Hansen (2009) and simulations in following sections show that the bias is non-negligible even with T as large as 20. As mentioned in the introduction, several bias-correction methods have been recently developed that could overcome this problem. Arellano and Hahn (2007) summarize the different approaches.

The methods can be grouped in three approaches based on the object that is corrected. The first one is to construct an analytical or numerical bias correction of a fixed effect estimator. Fernandez-Val (2009), among others, takes this approach to the problem and applies his analytical bias correction to dynamic binary choice models. The second group are those that correct the bias in moment equations. An example of this is Carro (2007) that uses an estimator of this type to correct the bias in dynamic binary choice models. The third group are those that correct the objective function. Arellano and Hahn (2006) and Bester and Hansen (2009) take this approach, with the latter including an application to a dynamic ordered probit model. Given that our model of interest is also a dynamic ordered probit, and that other alternatives will require some sort of transformation or derivations to be applied to our case, the HS-penalty estimator studied in Bester and Hansen (2009) is the first option we should consider. In addition to that, this estimator has the advantages of being simpler to compute than the Modified MLE in Carro (2007) and than the Bias Correction in Fernandez-Val (2009) because the HS does not require the calculation of expectations and the other two do. This advantage is more relevant in our case, because it has two fixed effects.

Arellano and Hahn (2007) shows the relations between the different type of approaches. Asymptotically all the methods and approaches are always reducing

the order of the bias of the MLE from the standard $O(T^{-1})$ to $O(T^{-2})$ for the general classes of models they were developed. However there may be differences when they are applied to specific cases . The following very simple example, used in Carro (2007), Arellano and Hahn (2007), and Bester and Hansen (2009), illustrates this point. Consider the model where $y_{it} \underset{iid}{\sim} N(\eta_i, \sigma_0^2)$. The ML estimator of σ_0^2 is $\widehat{\sigma}^2_{MLE} = \frac{1}{NT}\sum_i \sum_t (y_{it} - \widehat{\eta}_i)^2$. It is well known that $\widehat{\sigma}^2_{MLE}$ is not a consistent estimator of σ_0^2 when $N \to \infty$ with fixed T, since it converges to $\frac{T-1}{T}\sigma_0^2$. In this case the whole problem is very easy to fix. $\frac{1}{N(T-1)}\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it} - \widehat{\eta}_i)^2$ is the fixed T consistent estimator of σ_0^2. The MMLE from Carro (2007) produces this very same estimator, correcting not only the $O(T^{-1})$ term of the bias, but all the asymptotic bias in this special example. The HS removes the $O(T^{-1})$ term of the bias, but it does not attain the fixed-T consistent estimator. The one-step bias correction to the ML estimator from Fernandez-Val (2009) does not produce a fixed-T consistent estimator either, but its iterated form does. So, differences may appear between the different approaches when applied to specific models.

On the other hand, the incidental parameters problem can be seen as a finite sample bias problem in panel data context. The problem is not very important when T is large relative to N. However, since our panel does not have a large number of periods it is reasonable to wonder whether the good asymptotic properties of the MLE when T goes to infinity (sufficiently fast) are a good approximation to our finite sample. As a matter of fact, our problem is that the MLE has large biases when T is not very large; and having large N exacerbates the problem because the number of incidental parameters increases with N . It seems from simulations that we would need panels with a much larger number of time periods than those usually found in practice. The important implication of all this is that we have to look at the finite sample performance of the estimators for our model and sample sizes. In the methods considered here this is done through Monte Carlo experiments. Unfortunately, Bester and Hansen (2009) do not compare the finite sample properties of the method they use with others for the ordered probit case because many of the other methods will require some derivation to get the specific correction for this case. They, however, make such a comparison using a binary choice (probit and logit) models. Also, Carro (2007) and Fernandez-Val (2009) make Monte Carlo experiments for logit and probit models with different sample sizes (both in T and N), allowing us to compare a wide range of methods

for these models. From these comparisons we can conclude that the HS penalty approach is clearly not the best one and for sample sizes with T smaller than 13 the reminding bias can still be significant. Given this result, we should consider other of the proposed methods to estimate our ordered probit and evaluate its finite sample properties. Interesting candidates are the corrections discussed by Fernandez-Val (2009) and Carro (2007) since they are equally superior to other alternatives in finite sample performance in the relevant existing comparisons. In the next subsection we derive explicit formulas of the modified MLE used in Carro (2007) for the model considered here and evaluate its finite sample performance.

2.2. MMLE for a dynamic ordered probit with two fixed effects. The model to be estimated is defined in (7) and (8), and its log-likelihood is (11). Let $\gamma = (\beta, \rho_1, \rho_{-1})$ and $\eta_i = (\alpha_i, c_i)$. Partial derivatives will be denoted by the letter d, so the first order conditions will be $\mathbf{d}_{\eta i}(\gamma, \eta_i) \equiv \frac{\partial l_i(\gamma,\eta_i)}{\partial \eta_i}$ and $\mathbf{d}_{\gamma i}(\gamma, \eta_i) \equiv \frac{\partial l_i(\gamma,\eta_i)}{\partial \gamma}$. Bold letters represent vectors.

The MLE of η_i for given γ, $\eta_i(\gamma)$, solves $\mathbf{d}_{\eta i}(\gamma, \eta_i) = 0$. Then, the MLE of γ is obtained by maximizing the concentrated log-likelihood ($\sum_{i=1}^{N} l_i(\gamma, \eta_i(\gamma))$), i.e. by solving the following first order condition:

$$\frac{1}{T\,N} \sum_{i=1}^{N} \mathbf{d}_{\gamma i}(\gamma, \eta_i(\gamma)) = 0 \tag{12}$$

where $\mathbf{d}_{\gamma i}(\gamma, \eta_i(\gamma)) = \left. \frac{\partial l_i(\gamma,\eta_i)}{\partial \gamma} \right|_{\eta_i = \eta_i(\gamma)}$.

To reduce the bias of the estimation, we follow Carro (2007) in modifying the score of the concentrated log-likelihood adding a term that takes away the first order term of the asymptotic bias in T. By doing this, we get that the MMLE of the γ parameters of model (8) is the value that solves the following score equation:

$$\begin{aligned}
\mathbf{d}_{\gamma M i}(\gamma) = \mathbf{d}_{\gamma i}(\gamma, \eta_i(\gamma)) - \frac{1}{2} \frac{1}{{d_{\alpha\alpha i} d_{cci} - d_{\alpha c i}}^2} \Bigg[d_{\alpha\alpha i} \left(\mathbf{d}_{\gamma cci} + d_{\alpha cci} \frac{\partial \widehat{\alpha}_i}{\partial \gamma} + d_{ccci} \frac{\partial \hat{c}_i}{\partial \gamma} \right) \\
+ d_{cci} \left(\mathbf{d}_{\gamma\alpha\alpha i} + d_{\alpha\alpha\alpha i} \frac{\partial \widehat{\alpha}_i}{\partial \gamma} + d_{\alpha\alpha c i} \frac{\partial \hat{c}_i}{\partial \gamma} \right) - 2 d_{\alpha c i} \left(\mathbf{d}_{\gamma\alpha c i} + d_{\alpha\alpha c i} \frac{\partial \widehat{\alpha}_i}{\partial \gamma} + d_{\alpha c c i} \frac{\partial \hat{c}_i}{\partial \gamma} \right) \Bigg] \\
- \frac{\partial}{\partial \alpha_i} \left(\frac{E(\mathbf{d}_{\gamma c i}) E(d_{\alpha c i}) - E(d_{cci}) E(\mathbf{d}_{\gamma\alpha i})}{E(d_{\alpha\alpha i}) E(d_{cci}) - [E(d_{\alpha c i})]^2} \right) \Bigg|_{\eta_i = \eta_i(\gamma)} \\
- \frac{\partial}{\partial c_i} \left(\frac{E(\mathbf{d}_{\gamma\alpha i}) E(d_{\alpha c i}) - E(d_{\alpha\alpha i}) E(\mathbf{d}_{\gamma c i})}{E(d_{\alpha\alpha i}) E(d_{cci}) - [E(d_{\alpha c i})]^2} \right) \Bigg|_{\eta_i = \eta_i(\gamma)} = 0
\end{aligned} \tag{13}$$

where $\mathbf{d}_{\gamma i}(\gamma, \eta_i(\gamma))$ is the standard first order condition from the concentrated log-likelihood, as in (12). $\mathbf{d}_{\gamma ci} = \frac{\partial^2 l_i}{\partial \gamma \partial c_i}$, $d_{\alpha\alpha i} = \frac{\partial^2 l_i}{\partial \alpha_i^2}$, $\mathbf{d}_{\gamma\alpha ci} = \frac{\partial^3 l_i}{\partial \gamma \partial c_i \partial \alpha_i}$, and so on. $\widehat{\alpha}_i(\gamma)$ and $\widehat{c}_i(\gamma)$ are obtained from the first order conditions of α_i and c_i, as it is done in order to concentrate the log-likelihood. All expectations are conditional on the same set of information as the likelihood. These expectations can be computed by conditioning recursively, like we do to write the conditional likelihood. The parametric model (equations (7), (8) and the assumption about ε_{it}) from which we write the likelihood will also give the parametric form of the expectations we need to calculate.

We show in Appendix A how this modification on the score of the concentrated log-likelihood in (13) is a first order adjustment on the asymptotic bias of the ML score, so the first order condition is more nearly unbiased and the order of the bias of the estimator is reduced from $O(T^{-1})$ to $O(T^{-2})$.

2.3. Simulations. We simulate the model in equations (7), (8) with the following value of the parameters and Data Generating Process (DGP): $\beta = 1$, $\rho_1 = 0.5$, and $\rho_{-1} = -0.5$. The error follows a normal distribution: $\varepsilon_{it} \sim N(0, 1)$. The fixed effects are constructed as follows:

$$\alpha_i = \frac{1}{2} \sum_{t=1}^{4} x_{it} + u_i, \quad \text{where } u_i \sim N(x_{i0}, 1) \tag{14}$$

$$c_i = |z_i|, \quad \text{where } z_i \sim N(x_{i0}, 1). \tag{15}$$

so that they are correlated with the explanatory variables.[5] x_{it} follows a Gaussian AR(1) with autoregressive parameter equal to 0.5. Initial conditions are $x_{i0} \sim N(0, 1)$ and $h_{i0}^* = \alpha_i + \beta_0 x_{i0} + \varepsilon_{i0}$. We perform 1000 replications, with a population of $N = 250$ individuals. For each simulation we estimate the MLE, the MMLE given by equation (13) and the HS estimator defined in Bester and Hansen (2009). That is, the HS estimator is the value of the parameters that maximize the following penalized objective function:

$$\sum_{i=1}^{N} lk_i(\beta, \rho_1, \rho_{-1}, \alpha_i, c_i) - \sum_{i=1}^{N} \frac{1}{2} trace\left(\widehat{I}_{\alpha c_i}^{-1} \widehat{V}_{\alpha c_i}\right) - \frac{k}{2} \tag{16}$$

[5]Note that Bester and Hansen (2009) only consider in their simulations of an ordered probit the case where the fixed effects are independent of the covariates. Correlation of the unobserved heterogeneity with the covariates, as here, makes the problem more severe and the estimators may have worse performance. However, we consider this situation to be more realistic.

where lk_i is the log likelihood of i, $\widehat{I}_{\alpha c_i}$ is the sample information matrix for $e_i = (\alpha_i, c_i)'$, $\widehat{V}_{\alpha c_i}$ is a HAC estimator of $Var\left(\frac{1}{\sqrt{T}}\frac{\partial l_i}{\partial e_i}\right)$, and $k = \dim(e_i)$.

TABLE 1. Monte Carlo Results. Dynamic Ordered Probit parameters

Parameter	β		ρ_1		ρ_{-1}	
True value	1		0.5		-0.5	
Estimator	Mean Bias	RMSE	Mean Bias.	RMSE	Mean Bias	RMSE
			$T = 4$			
MLE	0.816	0.828	-0.474	0.516	0.551	0.586
HS	0.796	0.809	-0.392	0.443	0.467	0.509
MMLE	0.172	0.182	-0.254	0.282	0.280	0.305
			$T = 8$			
MLE	0.335	0.341	-0.188	0.216	0.189	0.216
HS	0.247	0.254	-0.115	0.153	0.119	0.154
MMLE	0.073	0.086	-0.062	0.108	0.067	0.109
			$T = 10$			
MLE	0.257	0.263	-0.145	0.171	0.154	0.179
HS	0.170	0.178	-0.083	0.119	0.093	0.127
MMLE	0.052	0.067	-0.036	0.086	0.050	0.093
			$T = 12$			
MLE	0.210	0.215	-0.127	0.152	0.127	0.151
HS	0.127	0.134	-0.072	0.106	0.074	0.106
MMLE	0.040	0.054	-0.030	0.079	0.036	0.081
			$T = 16$			
MLE	0.154	0.159	-0.093	0.118	0.096	0.119
HS	0.081	0.088	-0.048	0.083	0.054	0.085
MMLE	0.026	0.041	-0.017	0.068	0.022	0.069
			$T = 20$			
MLE	0.122	0.127	-0.072	0.095	0.078	0.101
HS	0.058	0.065	-0.034	0.067	0.042	0.074
MMLE	0.019	0.034	-0.009	0.058	0.016	0.062

Note: See a detailed description of the model simulated and other characteristics of the DGP in subsection 2.3.

Results from this experiment for different T are reported in Table 1, which shows the mean bias and the Root Mean Squared Error (RMSE). We find that for all T, the MMLE performs much better than the other two estimators. Comparing it with the HS, the differences are of greater magnitude for $T = 4$ and $T = 8$, where the HS is closer to the MLE than to the MMLE. When using the MMLE the bias is small than 10% of the true values with $T = 10$ for all but for one

of the ρ parameters. With $T = 12$ the bias when using the MMLE is already negligible whereas the HS contain biases and RMSE larger than the MMLE with $T = 10$. Even with $T = 16$ the HS exhibit mean biases greater than the MMLE with $T = 10$. It is not until $T = 20$ that the HS has small biases and RMSE. So HS needs a larger number of periods (at least larger than 16) to have small finite sample biases. Given this and the fact that the sample sizes we have in the empirical application of this paper are smaller than $T = 14$, we will use MMLE.

TABLE 2. Monte Carlo Results. Dynamic Ordered Probit parameters with different degrees of state dependence

Parameter	β		ρ_1		ρ_{-1}	
Estimator	Mean Bias	RMSE	Mean Bias.	RMSE	Mean Bias	RMSE
True value	1		-1		1	
MLE	0.204	0.212	-0.264	0.284	0.244	0.265
HS	0.105	0.116	-0.094	0.136	0.087	0.130
MMLE	0.012	0.044	-0.008	0.089	0.003	0.086
True value	1		-0.5		0.5	
MLE	0.212	0.218	-0.214	0.235	0.206	0.227
HS	0.116	0.126	-0.079	0.119	0.078	0.119
MMLE	0.026	0.048	-0.018	0.083	0.018	0.083
True value	1		0		0	
MLE	0.227	0.233	-0.180	0.201	0.180	0.201
HS	0.136	0.144	-0.079	0.116	0.084	0.119
MMLE	0.037	0.055	-0.028	0.082	0.032	0.084
True value	1		0.5		-0.5	
MLE	0.257	0.263	-0.145	0.171	0.154	0.179
HS	0.170	0.178	-0.083	0.119	0.093	0.127
MMLE	0.052	0.067	-0.036	0.086	0.050	0.093
True value	1		1		-1	
MLE	0.297	0.303	-0.105	0.144	0.111	0.148
HS	0.215	0.222	-0.086	0.126	0.091	0.129
MMLE	0.065	0.078	-0.057	0.100	0.069	0.107

Note: 1000 Monte Carlo simulations of the Ordered Probit model in equations (7) and (8), following the same DGP as in Table 1 (which is described at the begining of section 2.3), but changing the value of the state dependence parameters from negative to positive values, including the case with no state dependence.

To see if these results are maintained under different state dependence scenarios, we present in Table 2 simulations for different values of ρ_1 and ρ_{-1}, with

$T = 10$. The DGP is the same as that of Table 1, but the values of the state dependence parameters change from negative to positive, including the case with no state dependence. We find that the MMLE performs better than the other methods for all cases, in terms of bias and RMSE. In principle, having a more negative state dependence may improve all the estimators since it induces higher variance in y_{it}. This is the case for the estimation of β, where the three estimation methods improve, but it is not the case for the estimation of ρ_1 and ρ_{-1}, where the MMLE improves but the MLE and HS get worse.

TABLE 3. Monte Carlo Results. Inference over Dynamic Order Probit parameters: Conference intervals coverage and estimation of the estandard error.

Parameter	β		ρ_1		ρ_{-1}	
True value	1		0.5		-0.5	
	% Coverage		% Coverage		% Coverage	
Estimator	C.I. 95%	SE/SD	C.I. 95%	SE/SD	C.I. 95%	SE/SD
			$T = 8$			
MLE	0%	0.85	47%	0.87	48%	0.90
HS	0%	0.86	74%	0.91	73%	0.94
MMLE	64%	1.02	87%	0.93	85%	0.96
			$T = 10$			
MLE	0%	0.81	54%	0.91	53%	0.91
HS	3.5%	0.83	82%	0.96	78%	0.95
MMLE	74%	0.94	90%	0.96	89%	0.96
			$T = 12$			
MLE	0%	0.89	58%	0.91	62%	0.93
HS	8.8%	0.92	85%	0.96	83%	0.98
MMLE	81%	1.00	92%	0.95	92%	0.97
			$T = 16$			
MLE	0%	0.92	69%	0.91	68%	0.94
HS	29%	0.95	88%	0.96	88%	0.99
MMLE	88%	1.00	93%	0.94	93%	0.96
			$T = 20$			
MLE	2%	0.90	77%	0.96	73%	0.94
HS	48%	0.93	91%	1	88%	0.98
MMLE	90%	0.97	95%	0.98	93%	0.95

Note: We have used the inverse of the hessian as estimator of variance.

Finally, we consider the quality of inference based on these estimators. Table 3 present the coverage of 95% confidence intervals and the estimated asymptotic standard errors divided by the Standard Deviation. The latter is very close to 1 in all cases for the MMLE and in most cases for the other estimators, which indicates that de variance is being estimated well and the important problem is the bias. With respect to inference, the coverage of the confidence intervals is extremely poor for the MLE, specially for β. Even with $T = 20$, the coverage for β is smaller than 3%. The HS estimator improves inference with respect to the MLE, but it is still too far from the theoretical coverage of 95%, being also here the coverage for β specially bad even with $T = 20$. As it happens with the bias and RMSE criteria, the MMLE is clearly the best estimator of these three for doing inference, for all periods and parameters.

3. Empirical application: self-assessed health status in the British Household Panel

Self-assessed health (SAH) measures have been used as a proxy for true overall individual health status in many socioeconomic studies. Also, it has been shown to be a good predictor of mortality and of subsequent demand for medical care. This motivates the study of dynamics and potential explanatory factors of SAH. Moreover, SAH measures exhibit high persistence and it is interesting to know the relative contributions of state dependence and heterogeneity to it. In this section we estimate a dynamic ordered probit of SAH with two fixed effects, using MMLE whose properties has been studied in previous section.

Our model, in contrast with previous studies like Contoyannis, Jones, and Rice (2004), includes two fixed effects: one in the linear index equation and another one in the cut points. The motivation for doing this is to account for heterogeneity in reporting behavior (cut-points shifts) among individuals, in addition to accounting for unobserved factors that affect health status (index shift). The cut-point shifts occur if individuals use different thresholds when assessing their health and reporting it in the SAH categorical variable, so that they report a different value of SAH even though having the same level of true health. To control for these two, possibly correlated with other explanatory variables and between each other, unobserved factors, we include individual effects not only in the levels of the ordered probit but also in the cut points. Even though we cannot separately identified

these two sources of unobserved heterogeneity because we have to normalize one of the shifters, we are robustly controlling for both of them. In contrast, a model with unobserved heterogeneity in only one shifter and imposing homogeneity in all the other shifters will almost always give incorrect estimates if the two mentioned sources of heterogeneity are relevant.

The model we estimate is in equations (7) and (8), where $h_{it} = -1$ corresponds to the situation where poor health is reported, $h_{it} = 0$ to fair health, and $h_{it} = 1$ to good health. α_i and c_i are the model's fixed effects, and $\varepsilon_{it} \underset{iid}{\sim} N(0,1)$. The explanatory variables included in the model are described in the following subsection.

3.1. Data and variables. For our empirical analysis, we use the British Household Panel Survey (BHPS). This is a longitudinal survey of private households in Great Britain, and was designed as an annual survey of each adult (16+) member of a representative sample of more that 5,000 households, with a total of approximately 10,000 individual interviews. The same individuals are re-interviewed in successive waves and, if they split off from their original households are also re-interviewed along with all adult members of their new households. Similarly, new members joining sample households become eligible for interview and children are interviewed as they reach the age of 16. Currently, sixteen waves of data for the years 1991 - 2006 are available. We take into account individuals who gave a full interview at each wave. An unbalanced panel of individuals who were interviewed in at least 8 subsequent waves is used. Our sample consists of 76,128 observations from 6,375 individuals.

SAH is defined for waves 1-8 and 10-16 as the response to the question "Compared to people of your own age, would you say your health over the last 12 months on the whole has been: excellent, good, fair, poor, very poor?" At wave 9 the SAH question and categories were reworded. This makes the comparison with other waves difficult and wave 9 is not used in our empirical analysis.

The original five SAH categories were collapsed to a three-category variable, creating a new SAH variable, that will be our dependent variable, with the following codes: poor ($h_{it} = -1$) for individuals who reported either "very poor" or "poor" health; fair ($h_{it} = 0$) for individuals who reported "fair" health; and Good ($h_{it} = 1$) for individuals who reported "good" or "excellent" health.

The explanatory x variables in (7) can be grouped in three categories:

(1) Socioeconomic variables: three dummy variables representing marital status (Married, Widowed, Divorced/Separated), with Single as the reference category; five dummy variables representing employment (Self employed, In paid employment, Unemployed, Retired, Long term sick or disabled), with Other (Looking after family or home, On maternity leave, On a government training scheme, Full-time student/at school, Something else) as the reference category; size of the household (the number of people living in the same household); and number of kids in the household. The income variable is the logarithm of equivalised real income, adjusted using the Retail Price Index and equivalised by the McClement's scale to adjust for household size and composition, and consists on the sum of non-labour income and labour income in the reference year.

(2) Health variables: Among the explanatory variables of overall self-assessed heath status, we include information on objective health problems. The BHPS contains several questions about health problems and health care demand, but many of them can be induced by a self valuation that might differ from true health as much as SAH, and in an unobserved way. For example the number of visits to the doctor can be determined by a perception of a health problem rather than a true health problem. To avoid this endogeneity bias, we have selected only those questions that we regard as measuring more objective health situations and, therefore, are not affected by personal health assessments. We introduce the following variables:

- Health problems: This is a dummy variable, which takes the value 1 if the individual reports he/she has at least one of the following *permanent* health problems or disabilities: arthritis or rheumatism, difficulty in hearing, allergies, asthma, bronchitis, blood pressure, diabetes, migraine or frequent headaches, cancer and stroke, among others.

- Health limits daily activities: This is a dummy variable, which takes the value 1 if the individual answers 'yes' to the following question: does your health in any way limit your daily activities, compared to most people of your age? Examples of daily activities included are: doing the housework, climbing stairs, dressing yourself, walking for at least 10 minutes, etc.

- Health limits ability to work: Similar to previous question.
- Number of days in a Hospital as an in-patient in the reference year.

(3) Other controls: We include year dummies (excluding the necessary number to avoid prefect colinearity), age and age square. Note that the question about SAH that we use to construct our dependent variable asks for a comparison with the health of people with the same age as the respondent. However, there is a trend for SAH to become worse over time in the raw sample data that may indicate that the age effect over health is not being totally discounted by the respondents. This can be seen in table 5.[6] This is the reason for including age as explanatory variable.

TABLE 4. Number of individuals that reports each category of SAH by number of times it is reported.

Number	Excellent or good		Fair		Poor or very poor	
of times	Freq.	%	Freq. (N)	%	Freq. (N)	%
0	273	4.28	2076	32.56	4380	68.71
1	170	2.67	1114	17.47	898	14.09
2	182	2.85	867	13.60	367	5.76
3	193	3.03	641	10.05	213	3.34
4	233	3.65	481	7.55	137	2.15
5	273	4.28	376	5.90	99	1.55
6	379	5.95	279	4.38	79	1.24
7	456	7.15	204	3.20	46	0.72
8	665	10.43	145	2.27	47	0.74
9	563	8.83	83	1.30	33	0.52
10	533	8.36	61	0.96	32	0.50
11	495	7.76	19	0.30	16	0.25
12	544	8.53	20	0.31	8	0.13
13	672	10.54	5	0.08	9	0.14
14	744	11.67	4	0.06	11	0.17
Total	6375	100.00	6375	100.00	6375	100.00

Variables that are time-constant and specific for individuals, like the level of education and gender are not included in the set of explanatory variables since they can not be separately identified from the permanent unobserved heterogeneity. Therefore, the fixed effects account for these variables as well as for unobserved characterictics, and we cannot separate their effects. Sometimes this is seen as

[6]See Contoyannis, Jones, and Rice (2004) for further discussion on this.

TABLE 5. Proportion (in %) of each category of SAH by several characteristics

Characteristics and their Sample Proportions		SAH categories		
		Excellent or good	Fair	Poor or very poor
	All	73.19	19.39	7.42
By age group				
40.17	< 40	78.31	16.50	5.19
43.92	40-64	72.92	18.91	8.17
15.91	65+	61.02	28.02	10.96
By sex				
46.84	Male	75.35	18.32	6.34
53.16	Female	71.29	20.34	8.37
By marital status				
63.46	Married	74.00	18.86	7.14
8.92	Divorced	69.63	19.29	11.08
6.32	Widowed	58.84	28.92	12.25
21.3	Single	76.52	18.20	5.28
Health problems				
58.46	Yes	60.57	27.26	12.16
41.54	No	90.95	8.32	0.74
Health limits daily activities				
13.36	Yes	22.49	39.13	38.38
86.64	No	81.01	16.35	2.64
Health limits work				
16.43	Yes	29.85	38.29	31.86
83.57	No	81.71	15.68	2.61

TABLE 6. Sample transition probabilities from SAH in t-1 to SAH in t

		SAH in t			
		Excellent or good	Fair	Poor or very poor	Total
SAH	Excellent	85.91	11.84	2.25	100
in	Fair	43.22	45.18	11.59	100
$t-1$	Poor or very poor	17.66	31.60	50.74	100
	Proportion	72.80	19.67	7.53	100

a drawback of the fixed effects approach. However, the random effects approach only separately identifies the effect of these variables because of the unrealistic assumption that the unobserved characteristics are independent from them (for example that unobserved healthy life style is independent of education). Even with a correlated random effects approach, if correlation is allowed in a Mundlak (1978)

and Chamberlain (1984) style and initial conditions are controlled for following the proposal in Wooldridge (2005), it is not possible to separately identify the effect of these time constant variables from the effect of the unobserved factors correlated with them. For instance, Contoyannis, Jones, and Rice (2004) follows Wooldridge (2005) proposal and they comment about this impossibility of separating the effect of variables like education from the effect of the unobservables correlated with them.

Tables 4, 5 and 6 contain some descriptive numbers of the self-assessed heath reported in our sample. The most frequent category is excellent or good with more than 70% of the answers corresponding to this category. Also, there is high persistence in SAH reported, as can be seen in table 6, which shows the transition probabilities. In this table, the largest numbers are in the diagonal for all three values of SAH_{t-1}. Table 5 presents the variation on SAH across different characteristics and health variables. People that smoke tend to select worse self-assessed health categories than those that do not smoke. Married or single people respond the excellent or good health category more frequently than widows or divorced. The three objective health measures in table 5 alter the SAH responses in the expected direction and in greater magnitude than the socioeconomic variables also presented in the table.

Although there are clear connections, this empirical application does not substitute Contoyannis, Jones, and Rice (2004) since the latter contains a more detailed data description, makes further discussion of the estimated model and address other issues, like sample attrition, that are not considered in this paper.[7] However our paper complements Contoyannis, Jones, and Rice (2004) in several ways:

(i): We use more periods from the BHPS than them. They only use the first eight waves because the ninth contains a different question and categorization about SAH. While we drop the 9th wave too, we incorporate the waves after the 9th in our estimation. Since the model specified includes

[7]An unbalanced panel (with random attrition) in a dynamic panel model does not pose any complication to a fixed effect estimator (as opposed to a random effects estimator), as long as it does not imply many individuals with a very small number of periods; and in our sample all observations have at least 8 periods. The real problem here is that the assumption of attrition at random seems unrealistic. Contoyannis, Jones, and Rice (2004) made a test an find evidence of non-random attrition, but they also find that the bias this may be causing on the estimates seems to be negligible. Given that result and that this problem would take us too far from the main theme of this paper, we do not consider it here.

only one lag of h_{it}, we have all the variables we need for the 11th to 16th waves. For the 10th wave we have all the variables but h_{it-1} as it happens with the first wave. We treat the 10th wave like an initial observation and we condition it out in our likelihood leaving the probability of that observation totally unrestricted. Contoyannis, Jones, and Rice (2004) can not do this because of their way of solving the initial conditions problem and the use of random effects.

(ii): In our model we have two individual specific effects: one in the linear index and one in the cut points. Lindeboom and Van Doorslaer (2004) tests the existence of cut-point shifts and find clear evidence of different reporting behavior (cut-point shifting) for gender and age. Given that Contoyannis, Jones, and Rice (2004) are imposing homogeneous cut points, they estimate different models by gender to allow for that differing reporting behavior, but they do not allow unrestricted different behavior by age. Although we can not separately identify both sources of unobserved heterogeneity, our approach is robust to heterogenous cut points freely correlated with any of the determinants of SAH.

(iii): Use of fixed effects instead of random effects approach. The main advantages of this are that no arbitrary restriction is imposed in the correlation between the permanent unobserved heterogeneity and the observable variables, and that there is no initial conditions problem.

(iv): As an additional complement, our study includes some objective health measures, so we can see how much it is explained by the socioeconomic variables and by state dependence even after these measures are included.

3.2. Estimates. Table 7 presents the coefficient estimates for the dynamic ordered probit model based on three different estimators, that also includes different specification of the heterogeneity. The first estimated model (column I) is a pooled model without individual specific effects. The second (column II) is a correlated random effects specification with an individual effect in the linear index equation (the α_i parameter in (7)), but with homogeneous cut points. In this correlated random effects specification:

$$\alpha_i = \gamma_0 + \gamma_1' h_{i1} + \gamma_2' \overline{x}_i + u_i \tag{17}$$

TABLE 7. Estimates

Variables	I Pooled	II Correlated Random Effects	III MMLE
Health in t-1: Good	0.8366*** (0.0129)	0.3402*** (0.0231)	0.3696*** (0.0226)
Health in t-1: Poor	-0.5833*** (0.0206)	-0.3057*** (0.0334)	-0.2784*** (0.0296)
Age	0.0064*** (0.0023)	0.0004 (0.0208)	-0.0215(0.0282)
Age square	0.0000 (0.0000)	-0.0002** (0.0001)	-0.0003*** (0.0001)
Married	-0.0378** (0.0185)	0.0956 (0.0748)	0.0350(0.0672)
Separated/Divorced	-0.0954*** (0.0245)	0.1113 (0.1015)	0.0340(0.0817)
Widowed	-0.0683** (0.0292)	0.1902 (0.1330)	0.0474(0.1110)
Self employed	0.0858*** (0.0279)	0.0194 (0.0664)	0.0216(0.0660)
In paid employment	0.0344* (0.0198)	0.0586 (0.0418)	0.1069** (0.0425)
Unemployed	-0.0191 (0.0372)	0.0850 (0.0617)	0.0946(0.0680)
Retired	0.0101 (0.0294)	-0.0287 (0.0701)	0.1104* (0.0651)
Long term sick or disabled	-0.2478*** (0.0354)	-0.2552*** (0.0823)	-0.2562*** (0.0707)
Household size	-0.0422*** (0.0091)	0.0302 (0.0221)	-0.0127(0.0206)
Number kids	0.0434*** (0.0099)	0.0272 (0.0275)	0.0387* (0.0213)
Household Income	0.0571*** (0.0088)	-0.0116 (0.0194)	0.0112(0.0177)
Health problems	-0.6171*** (0.0138)	-0.6229*** (0.0279)	-0.7759*** (0.0334)
Health limits daily activities	-0.6421*** (0.0193)	-0.6147*** (0.0334)	-0.6865*** (0.0299)
Health limits work	-0.4297*** (0.0183)	-0.4243*** (0.0326)	-0.4854*** (0.0306)
Hospital days	-0.0306*** (0.0013)	-0.0371*** (0.0021)	-0.0350*** (0.0008)
Male	-0.0009 (0.0118)	0.0265 (0.0263)	
Non-white	-0.1028*** (0.0322)	-0.0496 (0.0708)	
Higher/1st degree	0.2312*** (0.0214)	0.2628*** (0.0470)	
HND/A level	0.1637*** (0.0168)	0.1983*** (0.0359)	
CSE/O level	0.1470*** (0.0153)	0.1890*** (0.0327)	
Cut point 1	-1.1358*** (0.0857)	-0.9473*** (0.2176)	
Cut point 2	0.1875*** (0.0855)	0.5676*** (0.2173)	
σ_u^2		0.3960	
Mean c_i			1.2775
Variance c_i			0.3942
Mean α_i			2.7760
Variance α_i			1.4170
N. obs.	69753	40140	16196
Log Lk.	-36943.35	-20304.28	-12198.37

Standard errors are reported in parenthesis. Estimates of year dummies in all models and within means of variables in random effects are not reported.
* significant at 10% ; ** significant at 5% ; *** significant at 1%.

where $\overline{x}_i$ is the average over the sample period of the exogenous variables, and $u_i \sim N(0, \sigma_u^2)$ independently of everything else. This is the kind of specification estimated in Contoyannis, Jones, and Rice (2004) that accounts for the correlated heterogeneity and the initial condition following Wooldridge (2005). The last specification (column III) is the specification described in previous subsections, that is the model in (7) and (8) treating α_i and c_i as fixed effects. It is estimated by MMLE.

To compare magnitudes of the effects across variables and estimates we will look at the relative effects (i.e. ratio of coefficients), and at the average and median marginal effects reported in tables 8 and 9 for the variables with a coefficient significantly different from zero.[8]

The pooled model exacerbates the state dependence effect due to the lack of permanent unobserved heterogeneity. Though not reported, we also estimated by MLE model in (7) and (8). As seen in the simulations it is severely biased, and that bias implies estimating much lower state depedance effects and higher effect of the other explanatory variables.

More interesting than that, is the comparison between the correlated random effects model and the fixed effects model (7) and (8) estimated by MMLE, columns II and III respectively of Tables 7, 8, and 9. The effect of all explanatory variables (with a significant effect) increases in absolute value in the MMLE case with respect to the random effects model. That includes also the state dependence effect (effect of h_{it-1}). Comparing coefficients in Table 7 we can also see that the effect of h_{it-1} increases proportionally less than the effect of the other relevant explanatory variables. In the Random effects specification the ratio of the coefficient of 'health problems' over the coefficient of $\mathbf{1}(h_{i,t-1} = good)$ is around 1.8, whereas in the MMLE that ratio is 2.1. In any case, this increase in the effect of state dependence,

[8]These marginal effects are also called partial effects. The marginal effects are averaged (or calculated their median) across the first eight waves of the panel as well as across the values of the covariates for each individual. This means that we first calculate the marginal effect for each individual in the sample at the observed values of the regressors and then we calculate the average (or the median) of them, instead of the marginal effect at the average value of the covariates. We do this in order to obtain summary measures of the marginal effects representative of the situation of the population (see Chamberlain, (1984))

TABLE 8. Average Marginal Effects on Probability of reporting good and poor health for significant variables.

(a) Good

	I		II		III	
	Pooled	St.Err.	Correlated Random Effects	St.Err.	MMLE	St.Err.
Health in t-1: Good	0.2400	0.0042	0.0762	0.0086	0.1122	0.0074
Health in t-1: Poor	-0.2012	0.0070	-0.0776	0.0101	-0.0832	0.0223
Age	0.0019	0.0001	-0.0043	0.0038	-0.0135	0.0080
Long term sick or disa.	-0.0608	0.0090	-0.0569	0.0196	-0.0729	0.0223
Health problems	-0.1434	0.0032	-0.1329	0.0146	-0.2277	0.0480
Health limits daily act.	-0.1760	0.0062	-0.1482	0.0136	-0.2045	0.0340
Health limits work	-0.1107	0.0053	-0.0977	0.0112	-0.1439	0.0141
Hospital days	-0.0068	0.0003	-0.0077	0.0008	-0.0099	0.0003

(b) Poor

	I		II		III	
	Pooled	St.Err.	Random Effects	St.Err.	MMLE	St.Err.
Health in t-1: Good	-0.0726	0.0017	-0.0198	0.0035	-0.0675	0.0877
Health in t-1: Poor	0.1091	0.0047	0.0241	0.0049	0.0650	0.0657
Age	-0.0007	0.0001	0.0014	0.0011	0.0088	0.0161
Long term sick or disa.	0.0225	0.0034	0.0168	0.0064	0.0547	0.0570
Health problems	0.0429	0.0010	0.0313	0.0054	0.1216	0.1667
Health limits daily act.	0.0649	0.0025	0.0431	0.0077	0.1501	0.1630
Health limits work	0.0389	0.0019	0.0273	0.0050	0.0994	0.1136
Hospital days	0.0024	0.0001	0.0022	0.0003	0.0065	0.0075

is remarkable because in the model in column III we are allowing for more, and more flexible permanent unobserved heterogeneity than in column II.[9]

Moreover, those differences in the estimated effects of the explanatory variables between the correlated random effects model and the fixed effects model estimated by MMLE are statistically significant. As it is known, if the restrictions imposed by the correlated random effects model are correct its estimates are far more precise (i.e. efficient) than the estimates of the fixed effects model (even after the modification of the MLE), thought both are consistent. Given this, we have made

[9]Remember here that permanent unobserved heterogeneity, state dependence and persistence in observable variables are alternative explanations of the observed high persistence in h_{it}.

TABLE 9. Median Marginal Effects on Probability of reporting good and poor health for significant variables.

(a) Good

	I Pooled	II Corr. Random Effects	III MMLE
Health in t-1: Good	0.2439	0.0771	0.1175
Health in t-1: Poor	-0.2140	-0.0802	-0.0898
Age	0.0019	-0.0037	-0.0130
Long term sick or disabled	-0.0615	-0.0580	-0.0781
Health problems	-0.1387	-0.1267	-0.2402
Health limits daily activities	-0.1806	-0.1525	-0.2193
Health limits work	-0.1130	-0.1001	-0.1540
Hospital days	-0.0069	-0.0077	-0.0105

(b) Poor

	I Pooled	II Random Effects	III MMLE
Health in t-1: Good	-0.0534	-0.0091	-0.0631
Health in t-1: Poor	0.1021	0.0143	0.0647
Age	-0.0003	0.0004	0.0080
Long term sick or disabled	0.0125	0.0075	0.0540
Health problems	0.0201	0.0108	0.1078
Health limits daily activities	0.0434	0.0232	0.1509
Health limits work	0.0234	0.0133	0.0976
Hospital days	0.0012	0.0008	0.0062

a Hausman type test to see if those important differences are only due to the more imprecise estimates in columns III. We have made the test over the Average Marginal effects in table 8 instead of the parameters in table 7 for two reasons. Firstly, Marginal Effects (including their average), and not the parameters in equations (7) and (8), are usually the actual parameters of interest in nonlinear models. Secondly, the average marginal effects do not suffer the different scales problem that makes magnitudes in columns II and III of Table 7 not directly comparable and not directly interpretable. The average marginal effects of both models are well defined within the same scale, as any other marginal effect over choice probabilities, and their magnitude has the same clear interpretation. If we were primarily

interested in a single average marginal effect, like the effect of $h_{i,t-1} = good$ over the probability of $h_{i,t} = good$, we can use a t-statistic that ignores the others. Doing this for all the average marginal effects there are four variables for which we reject, at 5%, the null hypothesis that both estimates are the same. Doing a joint test we also reject the null hypotesis that the correlated random effects estimates and the fixed effects MMLE estimates are the same, rejecting, therefore, the restriction imposed in the correlated random effects model.[10]

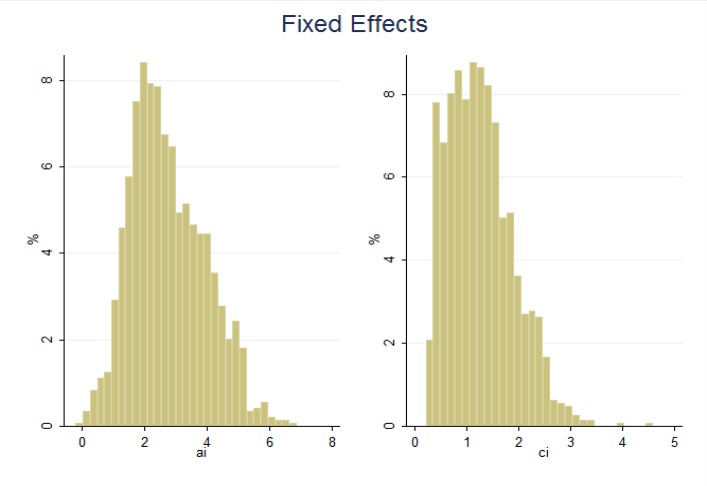


FIGURE 1. Distribution (histogram) of the fixed effects from MML estimates.

The previous two paragraphs are a clear indication that ignoring the added dimension of heterogeneity and the flexibility in the distribution of the fixed effects matters when estimating the model and the marginal effects of variables. It is not only a matter of the amount of heterogeneity but also a matter of the other restrictions being imposed in the model in column II.

Besides the formal test of random effects vs. fixed effects done, we look at the unobserved heterogeneity both in the linear index equation and in the cut point shift. Figure 1 displays the estimated distribution (histogram) of both fixed effects in the population. Both exhibit important variation. The average for α_i is 2.78 and 1.28 for c_i. The standard deviations of those distributions are 1.19 and 0.63

[10]In the Hausman test we have used the Var-Cov of the Fixed Effects estimates only, instead of subtracting from it the Var-Cov of the Random Effects. We do this in order to avoid the difference not being a positive definite matrix due to the use of different estimates of the variance of the errors. This represents a lower bound for this test and a rejection here will also be a rejection when using the well defined difference in the var-cov matrices.

respectively. α_i in the random effects specification is the compound equation (17) that includes a linear relation to some observables and an additive unobserved term that is assumed to follow a normal distribution. Given the estimates of the parameters of equation (17), the estimated average for α_i in the random effects model is 2.06, and its standard deviation is 0.64. With respect to the heterogeneity on the cut points, in (8) the second cut point has been normalized to be zero. Interestingly its estimate in the random effects model is not significantly different from zero. The average of $-c_i$, the first cut point, is very close to the estimate of the first cut point in the random effects specification. However, as can be seen in the right panel of figure 1 there is important variation in c_i among individuals that is ignored by the random effects model estimated.

TABLE 10. Proportion of individuals with marginal effects (on the probability of reporting good and poor) that are significantly different from zero at 10%.

Variable	Proportion	
	Good	Poor
Health in t-1: Good	50.38	11.06
Health in t-1: Poor	49.83	20.32
Age	30.20	4.31
Long term sick or disabled	43.49	17.33
Health problems	47.04	9.95
Health limits daily activities	49.13	19.69
Health limits work	48.50	17.61
Hospital days	47.11	15.52

Focusing on the MML estimates, the two indicators of h_{it-1} and the variables that capture objective health problems have a significant effect over SAH, with the expected signs. The apparently surprising result that 'Long term sick or disable' has much smaller effect than 'Health problems', 'Health limits daily activities', or 'Health limits work', is explained by the positive correlation between being 'long term sick or disable' and the other three. This means that limiting daily activities or work affects more negatively SAH than only suffering a long term sickness that does not limit us, although both situations are correlated.[11] It also means that for many people the effect of a disability or long term sickness is going to be the

[11]Sample correlation between 'Long term sick or disable' and 'Health limits daily activities' or 'Health limits work' is around 0.35.

composite effect of the 'Long term sick or disable' variable plus the effect of the limit in daily activities and work that it causes. As in Contoyannis, Jones, and Rice (2004) we also find evidence of strong positive state dependence, even after including more heterogeneity and the objective health measures. Apart from age and smoking, no socioeconomic variable has a significant effect. This is in contrast with the apparent correlation in the sample between these variables and SAH described in table 5.

In addition to looking at the average and median marginal effects reported in tables 8 and 9, we look at how many individuals have a significant marginal effect in the sample given their particular situation and unobserved characteristics. Table 10 presents the proportion of individuals with a significant (at 10%) marginal effects over the probability of reporting good and bad health, for the same variables as in table 8. Notice that although the average marginal effects are significant, there is a great deal of heterogeneity so that for half of the population the marginal effects over the probability of reporting good health is not significantly different from zero for many of these variables.

4. Conclusion

In this paper we have considered the estimation of a dynamic ordered probit with fixed effects of a self-assessed health status, which includes two fixed effect: one in the linear index equation and one in the cut points. These two fixed effects, instead of only one as usually done, are implied by the potential existence of two sources of heterogeneity: unobserved health status and reporting behavior. Heterogeneous reporting behavior means that individuals use different thresholds when assessing their health and reporting it, so that they report a different value of SAH even if the have the same true health. Even though we cannot separately identify these two sources of heterogeneity we are robustly controlling for them by using two fixed effects. Based on our best estimates, the two fixed effects exhibit important variation and it is relevant to account for both when estimating the effect of other variables. Our estimates also show that state dependence is very important even though we have controlled for unobserved heterogeneity and some forms of objective health measures. The latter are the variables with higher marginal effects.

The recent literature in bias-adjusted methods of estimation of nonlinear panel data models with fixed effects has produced several potentially equivalent estimators. Here we find that the most directly and easily applicable correction to our model, which is the HS estimator proposed in Bester and Hansen (2009), has still important biases in our sample size. This lead us to consider the Modified MLE proposed in Carro (2007). We derive the expression of the MMLE in our case, and perform Monte Carlo experiments to evaluate its finite sample properties and compare it with the HS. The MMLE has a negligible bias in our sample size. These Monte Carlo experiments contribute to the mentioned literature on bias-adjusted methods of estimation by showing how well two of the proposed methods work for a specific model and sample size. Also, this will be useful information for other applications when having to choose among the several correction methods.

Appendix A: Reduction of the order of the bias

In this appendix we show that the modified score presented above correct the first order asymptotic bias of the original score. This is done by deriving the leading term of the bias of the MLE's score, and then by showing that the modification is substracting that term from the score. This follows Carro (2007), adapting it to our model with two fixed effects.

The notation used is the same as before: we denote partial derivatives by the letter d; bold letters are used to denote vectors; the derivatives evaluated at the true values of the parameters are represented by including a 0 in the sub-index (e.g. $d_{\eta i0} = d_{\eta i}(\gamma_0, \eta_{i0})$).

4.1. Deriving the leading term of the bias of the score in the MLE. We start by deriving the first term of the bias in the score of the original unmodified concentrated log-likelihood. Expanding this score around η_{i0}, and evaluating it at γ_0 we get:

$$
\begin{aligned}
\mathbf{d}_{\gamma i}(\gamma_0, \eta_i(\gamma_0)) = \mathbf{d}_{\gamma i0} + \; & d_{\gamma ai0}(\hat{a}_i(\gamma_0) - a_{i0}) \\
& + \mathbf{d}_{\gamma ci0}(\hat{c}_i(\gamma_0) - c_{i0}) \\
& + \frac{1}{2}\mathbf{d}_{\gamma aai0}(\hat{a}_i(\gamma_0) - a_{i0})^2 + \frac{1}{2}\mathbf{d}_{\gamma cci0}(\hat{c}_i(\gamma_0) - c_{i0})^2 \\
& + \mathbf{d}_{\gamma aci0}(\hat{a}_i(\gamma_0) - a_{i0})(\hat{c}_i(\gamma_0) - c_{i0}) + O_p(T^{-1/2}) + \ldots
\end{aligned} \tag{18}
$$

Now we need expressions for $(\hat{a}_i(\gamma_0) - a_{i0})$ and $(\hat{c}_i(\gamma_0) - c_{i0})$, for which we do asymptotic expansions, following Rilstone, Srivastava and Ullah (1996):

$$(\hat{a}_i(\gamma_0) - a_{i0}) = b^a_{-1/2} + b^a_{-1} + O_p(T^{-3/2}) \tag{19}$$

$$(\hat{c}_i(\gamma_0) - c_{i0}) = b^c_{-1/2} + b^c_{-1} + O_p(T^{-3/2}) \tag{20}$$

where

$$b^a_{-1/2} = \frac{\frac{1}{T}d_{ci0}E\left(\frac{1}{T}\,d_{aci0}\right) - \frac{1}{T}d_{ai0}E\left(\frac{1}{T}\,d_{cci0}\right)}{E\left(\frac{1}{T}\,d_{aai0}\right)E\left(\frac{1}{T}\,d_{cci0}\right) - E\left(\frac{1}{T}\,d_{aci0}\right)^2} \tag{21}$$

$$b^c_{-1/2} = \frac{\frac{1}{T}d_{ai0}E\left(\frac{1}{T}\,d_{aci0}\right) - \frac{1}{T}d_{ci0}E\left(\frac{1}{T}\,d_{aai0}\right)}{E\left(\frac{1}{T}\,d_{aai0}\right)E\left(\frac{1}{T}\,d_{cci0}\right) - E\left(\frac{1}{T}\,d_{aci0}\right)^2} \tag{22}$$

It is also useful to obtain:

$$(\hat{a}_i(\gamma_0) - a_{i0})^2 = (b^a_{-1/2})^2 + O_p(T^{-3/2}) \tag{23}$$

$$(\hat{c}_i(\gamma_0) - c_{i0})^2 = (b^c_{-1/2})^2 + O_p(T^{-3/2}) \tag{24}$$

$$(\hat{a}_i(\gamma_0) - a_{i0})\,(\hat{c}_i(\gamma_0) - c_{i0}) = b^a_{-1/2}\, b^c_{-1/2} + O_p(T^{-3/2}) \tag{25}$$

With respect to the squares of $b^a_{-1/2}$ and $b^c_{-1/2}$, we get:

$$(b^a_{-1/2})^2 = \frac{\left(\frac{1}{T}d_{ai0}\right)^2 E\left(\frac{1}{T}\,d_{cci0}\right)^2 + \left(\frac{1}{T}d_{ci0}\right)^2 E\left(\frac{1}{T}\,d_{aci0}\right)^2 - 2\,\frac{1}{T}d_{ai0}\frac{1}{T}d_{ci0}E\left(\frac{1}{T}\,d_{aci0}\right)E\left(\frac{1}{T}\,d_{cci0}\right)}{\left(E\left(\frac{1}{T}\,d_{aai0}\right)E\left(\frac{1}{T}\,d_{cci0}\right) - E\left(\frac{1}{T}\,d_{aci0}\right)^2\right)^2}$$

$$(b^c_{-1/2})^2 = \frac{\left(\frac{1}{T}d_{ci0}\right)^2 E\left(\frac{1}{T}\,d_{aai0}\right)^2 + \left(\frac{1}{T}d_{ai0}\right)^2 E\left(\frac{1}{T}\,d_{aci0}\right)^2 - 2\,\frac{1}{T}d_{ai0}\frac{1}{T}d_{ci0}E\left(\frac{1}{T}\,d_{aai0}\right)E\left(\frac{1}{T}\,d_{aci0}\right)}{\left(E\left(\frac{1}{T}\,d_{aai0}\right)E\left(\frac{1}{T}\,d_{cci0}\right) - E\left(\frac{1}{T}\,d_{aci0}\right)^2\right)^2}$$

Substituting by expectations, and using the information matrix identity ($E(d_{aci}) = -E(d_{ai}d_{ci})$), we get:

$$(b^a_{-1/2})^2 = -\frac{1}{T}\,\frac{E\left(\frac{1}{T}\,d_{cci0}\right)}{E\left(\frac{1}{T}\,d_{aai0}\right)E\left(\frac{1}{T}\,d_{cci0}\right) - E\left(\frac{1}{T}\,d_{aci0}\right)^2} + O_p(T^{-3/2}) \tag{26}$$

$$(b^c_{-1/2})^2 = -\frac{1}{T}\,\frac{E\left(\frac{1}{T}\,d_{aai0}\right)}{E\left(\frac{1}{T}\,d_{aai0}\right)E\left(\frac{1}{T}\,d_{cci0}\right) - E\left(\frac{1}{T}\,d_{aci0}\right)^2} + O_p(T^{-3/2}) \tag{27}$$

Following the same procedure for the cross-product, we get:

$$b^a_{-1/2}\, b^c_{-1/2} = \frac{1}{T}\,\frac{E\left(\frac{1}{T}\,d_{aci0}\right)}{E\left(\frac{1}{T}\,d_{aai0}\right)E\left(\frac{1}{T}\,d_{cci0}\right) - E\left(\frac{1}{T}\,d_{aci0}\right)^2} + O_p(T^{-3/2}) \tag{28}$$

With respect to b^a_{-1} and b^c_{-1}, we follow the same procedure (replace by expectations and use the information matrix identity) to get:

(29)

$$
\begin{aligned}
b^a_{-1} = \frac{1}{2\,T} & \frac{1}{\left(E\left(\frac{1}{T}\,d_{aai0}\right)E\left(\frac{1}{T}\,d_{cci0}\right) - E\left(\frac{1}{T}\,d_{aci0}\right)^2\right)^2} \\
& \left\{ 2\,E\left(\frac{1}{T}\,d_{aci0}\right)^2 \left(E\left(\frac{1}{T}\,d_{acci0}\right) + E\left(\frac{1}{T}\,d_{ai0}d_{cci0}\right) + E\left(\frac{1}{T}\,d_{ci0}d_{aci0}\right)\right)\right. \\
& + E\left(\frac{1}{T}\,d_{cci0}\right)^2 \left[E\left(\frac{1}{T}\,d_{aaai0}\right) + 2E\left(\frac{1}{T}\,d_{ai0}d_{aai0}\right)\right] \\
& + E\left(\frac{1}{T}\,d_{aai0}\right)E\left(\frac{1}{T}\,d_{cci0}\right)\left[E\left(\frac{1}{T}\,d_{acci0}\right) + 2E\left(\frac{1}{T}\,d_{ci0}d_{aci0}\right)\right] \\
& - E\left(\frac{1}{T}\,d_{aci0}\right)E\left(\frac{1}{T}\,d_{aai0}\right)\left[E\left(\frac{1}{T}\,d_{ccci0}\right) + 2E\left(\frac{1}{T}\,d_{ci0}d_{cci0}\right)\right] \\
& \left. -E\left(\frac{1}{T}\,d_{aci0}\right)E\left(\frac{1}{T}\,d_{cci0}\right)\left[3E\left(\frac{1}{T}\,d_{aaci0}\right) + 4E\left(\frac{1}{T}\,d_{ai0}d_{aci0}\right) + 2E\left(\frac{1}{T}\,d_{ci0}d_{aai0}\right)\right]\right\} \\
& + O_p(T^{-3/2})
\end{aligned}
$$

(30)

$$
\begin{aligned}
b^c_{-1} = \frac{1}{2\,T} & \frac{1}{\left(E\left(\frac{1}{T}\,d_{aai0}\right)E\left(\frac{1}{T}\,d_{cci0}\right) - E\left(\frac{1}{T}\,d_{aci0}\right)^2\right)^2} \\
& \left\{ 2\,E\left(\frac{1}{T}\,d_{aci0}\right)^2 \left[E\left(\frac{1}{T}\,d_{aaci0}\right) + E\left(\frac{1}{T}\,d_{ci0}d_{aai0}\right) + E\left(\frac{1}{T}\,d_{ai0}d_{aci0}\right)\right]\right. \\
& + E\left(\frac{1}{T}\,d_{aai0}\right)^2 \left[E\left(\frac{1}{T}\,d_{ccci0}\right) + 2E\left(\frac{1}{T}\,d_{ci0}d_{cci0}\right)\right] \\
& + E\left(\frac{1}{T}\,d_{aai0}\right)E\left(\frac{1}{T}\,d_{cci0}\right)\left[E\left(\frac{1}{T}\,d_{aaci0}\right) + 2E\left(\frac{1}{T}\,d_{ai0}d_{aci0}\right)\right] \\
& - E\left(\frac{1}{T}\,d_{aci0}\right)E\left(\frac{1}{T}\,d_{cci0}\right)\left[E\left(\frac{1}{T}\,d_{aaai0}\right) + 2E\left(\frac{1}{T}\,d_{ai0}d_{aai0}\right)\right] \\
& \left. -E\left(\frac{1}{T}\,d_{aci0}\right)E\left(\frac{1}{T}\,d_{aai0}\right)\left[3E\left(\frac{1}{T}\,d_{acci0}\right) + 4E\left(\frac{1}{T}\,d_{ci0}d_{aci0}\right) + 2E\left(\frac{1}{T}\,d_{ai0}d_{cci0}\right)\right]\right\}
\end{aligned}
$$

(31)

$$
+ O_p(T^{-3/2})
$$

Introducing all these expressions in (18), and taking expectations, we get:

(32)

$$
\begin{aligned}
&E(d_{\gamma i}(\gamma_0, \hat{\eta}_i(\gamma_0))) = \\
&\frac{E\left(\frac{1}{T}\,\mathbf{d}_{\gamma ai0} d_{ci0}\right) E\left(\frac{1}{T}\,d_{aci0}\right) - E\left(\frac{1}{T}\,\mathbf{d}_{\gamma ai0} d_{ai0}\right) E\left(\frac{1}{T}\,d_{cci0}\right)}{E\left(\frac{1}{T}\,d_{aai0}\right) E\left(\frac{1}{T}\,d_{cci0}\right) - E\left(\frac{1}{T}\,d_{aci0}\right)^2} \\
&+ \frac{1}{2}\,\frac{E\left(\frac{1}{T}\,\mathbf{d}_{\gamma ai0}\right)}{\left(E\left(\frac{1}{T}\,d_{aai0}\right) E\left(\frac{1}{T}\,d_{cci0}\right) - E\left(\frac{1}{T}\,d_{aci0}\right)^2\right)^2} \\
&\left\{ 2\,E\left(\frac{1}{T}\,d_{aci0}\right)^2 \left(E\left(\frac{1}{T}\,d_{acci0}\right) + E\left(\frac{1}{T}\,d_{ai0} d_{cci0}\right) + E\left(\frac{1}{T}\,d_{ci0} d_{aci0}\right)\right)\right. \\
&+ E\left(\frac{1}{T}\,d_{cci0}\right)^2 \left[E\left(\frac{1}{T}\,d_{aaai0}\right) + 2E\left(\frac{1}{T}\,d_{ai0} d_{aai0}\right)\right] \\
&+ E\left(\frac{1}{T}\,d_{aai0}\right) E\left(\frac{1}{T}\,d_{cci0}\right) \left[E\left(\frac{1}{T}\,d_{acci0}\right) + 2E\left(\frac{1}{T}\,d_{ci0} d_{aci0}\right)\right] \\
&- E\left(\frac{1}{T}\,d_{aci0}\right) E\left(\frac{1}{T}\,d_{aai0}\right) \left[E\left(\frac{1}{T}\,d_{ccci0}\right) + 2E\left(\frac{1}{T}\,d_{ci0} d_{cci0}\right)\right] \\
&\left. -E\left(\frac{1}{T}\,d_{aci0}\right) E\left(\frac{1}{T}\,d_{cci0}\right) \left[3E\left(\frac{1}{T}\,d_{aaci0}\right) + 4E\left(\frac{1}{T}\,d_{ai0} d_{aci0}\right) + 2E\left(\frac{1}{T}\,d_{ci0} d_{aai0}\right)\right]\right\} \\
&+ \frac{E\left(\frac{1}{T}\,\mathbf{d}_{\gamma ci0} d_{ai0}\right) E\left(\frac{1}{T}\,d_{aci0}\right) - E\left(\frac{1}{T}\,\mathbf{d}_{\gamma ci0} d_{ci0}\right) E\left(\frac{1}{T}\,d_{aai0}\right)}{E\left(\frac{1}{T}\,d_{aai0}\right) E\left(\frac{1}{T}\,d_{cci0}\right) - E\left(\frac{1}{T}\,d_{aci0}\right)^2} \\
&+ \frac{1}{2}\,\frac{E\left(\frac{1}{T}\,\mathbf{d}_{\gamma ci0}\right)}{\left(E\left(\frac{1}{T}\,d_{aai0}\right) E\left(\frac{1}{T}\,d_{cci0}\right) - E\left(\frac{1}{T}\,d_{aci0}\right)^2\right)^2} \\
&\left\{ 2\,E\left(\frac{1}{T}\,d_{aci0}\right)^2 \left[E\left(\frac{1}{T}\,d_{aaci0}\right) + E\left(\frac{1}{T}\,d_{ci0} d_{aai0}\right) + E\left(\frac{1}{T}\,d_{ai0} d_{aci0}\right)\right]\right. \\
&+ E\left(\frac{1}{T}\,d_{aai0}\right)^2 \left[E\left(\frac{1}{T}\,d_{ccci0}\right) + 2E\left(\frac{1}{T}\,d_{ci0} d_{cci0}\right)\right] \\
&+ E\left(\frac{1}{T}\,d_{aai0}\right) E\left(\frac{1}{T}\,d_{cci0}\right) \left[E\left(\frac{1}{T}\,d_{aaci0}\right) + 2E\left(\frac{1}{T}\,d_{ai0} d_{aci0}\right)\right] \\
&- E\left(\frac{1}{T}\,d_{aci0}\right) E\left(\frac{1}{T}\,d_{cci0}\right) \left[E\left(\frac{1}{T}\,d_{aaai0}\right) + 2E\left(\frac{1}{T}\,d_{ai0} d_{aai0}\right)\right] \\
&\left. -E\left(\frac{1}{T}\,d_{aci0}\right) E\left(\frac{1}{T}\,d_{aai0}\right) \left[3E\left(\frac{1}{T}\,d_{acci0}\right) + 4E\left(\frac{1}{T}\,d_{ci0} d_{aci0}\right) + 2E\left(\frac{1}{T}\,d_{ai0} d_{cci0}\right)\right]\right\} \\
&+ \frac{1}{E\left(\frac{1}{T}\,d_{aai0}\right) E\left(\frac{1}{T}\,d_{cci0}\right) - E\left(\frac{1}{T}\,d_{aci0}\right)^2}\,{}^{84} \\
&\left[E\left(\frac{1}{T}\,\mathbf{d}_{\gamma aci0}\right) E\left(\frac{1}{T}\,d_{aci0}\right) - \frac{1}{2}E\left(\frac{1}{T}\,\mathbf{d}_{\gamma aai0}\right) E\left(\frac{1}{T}\,d_{cci0}\right) - \frac{1}{2}E\left(\frac{1}{T}\,\mathbf{d}_{\gamma cci0}\right) E\left(\frac{1}{T}\,d_{aai0}\right)\right] \\
&+ O(T^{-1})
\end{aligned}
$$

The remainder of this expression is $O(T^{-1})$ because $O_p(T^{-1/2})$ terms have zero mean. This means that the score of the original concentrated likelihood has a bias of order $O(1)$, whose expression is in the previous formulae.

4.2. Modified Score. The modified score in (13) can be decomposed in three terms, $\mathbf{d}_{\gamma Mi}(\gamma) = A + B + C$, such that:

$$A = \mathbf{d}_{\gamma i}(\gamma, \eta_i(\gamma)) \tag{33}$$

$$\begin{aligned} B = -\frac{1}{2} & \frac{1}{d_{aai}d_{cci} - {d_{aci}}^2} \\ & \left[d_{aai} \left(\mathbf{d}_{\gamma cci} + d_{acci} \frac{\partial \hat{a}_i}{\partial \gamma} + d_{ccci} \frac{\partial \hat{c}_i}{\partial \gamma} \right) \right. \\ & + d_{cci} \left(\mathbf{d}_{\gamma aai} + d_{aaai} \frac{\partial \hat{a}_i}{\partial \gamma} + d_{aaci} \frac{\partial \hat{c}_i}{\partial \gamma} \right) \\ & \left. -2d_{aci} \left(\mathbf{d}_{\gamma aci} + d_{aaci} \frac{\partial \hat{a}_i}{\partial \gamma} + d_{acci} \frac{\partial \hat{c}_i}{\partial \gamma} \right) \right] \end{aligned} \tag{34}$$

$$\begin{aligned} C = & -\frac{\partial}{\partial a_i} \left(\frac{E(\mathbf{d}_{\gamma ci})E(d_{aci}) - E(d_{cci})E(\mathbf{d}_{\gamma ai})}{E(d_{aai})E(d_{cci}) - [E(d_{aci})]^2} \right) \Bigg|_{\eta_i = \eta_i(\gamma)} \\ & -\frac{\partial}{\partial c_i} \left(\frac{E(\mathbf{d}_{\gamma ai})E(d_{aci}) - E(d_{aai})E(\mathbf{d}_{\gamma ci})}{E(d_{aai})E(d_{cci}) - [E(d_{aci})]^2} \right) \Bigg|_{\eta_i = \eta_i(\gamma)} \end{aligned} \tag{35}$$

A is the score of the original un-modified concentrated log-likelihood. So, we now analyze B and C.

Part B. We first want to derive expression for $\partial \hat{a}_i / \partial \gamma$ and $\partial \hat{c}_i / \partial \gamma$. Differentiating the score of the concentrated log-likelihood, $\mathbf{d}_{\eta i}(\gamma, \eta_i(\gamma))$, with respect to γ we get a system of two equations with two unknowns. Solving for $\partial \hat{a}_i / \partial \gamma$ and $\partial \hat{c}_i / \partial \gamma$ we get:

$$\frac{\partial \hat{a}_i(\gamma)}{\partial \gamma} = \frac{\mathbf{d}_{\gamma ci} d_{aci} - d_{cci} \mathbf{d}_{\gamma ai}}{d_{aai} d_{cci} - d_{aci}^2} \tag{36}$$

$$\frac{\partial \hat{c}_i(\gamma)}{\partial \gamma} = \frac{\mathbf{d}_{\gamma ai} d_{aci} - d_{aai} \mathbf{d}_{\gamma ci}}{d_{aai} d_{cci} - d_{aci}^2} \tag{37}$$

evaluating at γ_0 and replacing by expectations:

$$\frac{\partial \hat{a}_i(\gamma_0)}{\partial \gamma} = \frac{E\left(\frac{1}{T}\,\mathbf{d}_{\gamma ci0}\right) E\left(\frac{1}{T}\,d_{aci0}\right) - E\left(\frac{1}{T}\,d_{cci0}\right) E\left(\frac{1}{T}\,\mathbf{d}_{\gamma ai0}\right)}{E\left(\frac{1}{T}\,d_{aai0}\right) E\left(\frac{1}{T}\,d_{cci0}\right) - E\left(\frac{1}{T}\,d_{aci0}\right)^2} + O_p(T^{-\frac{1}{2}}) \tag{38}$$

$$\frac{\partial \hat{c}_i(\gamma_0)}{\partial \gamma} = \frac{E\left(\frac{1}{T}\,\mathbf{d}_{\gamma ai0}\right) E\left(\frac{1}{T}\,d_{aci0}\right) - E\left(\frac{1}{T}\,d_{aai0}\right) E\left(\frac{1}{T}\,\mathbf{d}_{\gamma ci0}\right)}{E\left(\frac{1}{T}\,d_{aai0}\right) E\left(\frac{1}{T}\,d_{cci0}\right) - E\left(\frac{1}{T}\,d_{aci0}\right)^2} + O_p(T^{-\frac{1}{2}}) \tag{39}$$

Introducing in (34) and rearranging terms:

$$\begin{aligned} B = & -\frac{E\left(\frac{1}{T}\,\mathbf{d}_{\gamma ci0}\right) E\left(\frac{1}{T}\,d_{aci0}\right) - E\left(\frac{1}{T}\,d_{cci0}\right) E\left(\frac{1}{T}\,\mathbf{d}_{\gamma ai0}\right)}{E\left(\frac{1}{T}\,d_{aai0}\right) E\left(\frac{1}{T}\,d_{cci0}\right) - E\left(\frac{1}{T}\,d_{aci0}\right)^2} \\ & \frac{d_{aai}d_{acci} + d_{cci}d_{aaai} - 2d_{aci}d_{aaci}}{2(d_{aai}d_{cci} - d_{aci}^2)} \\ & - \frac{d_{aai}d_{acci} + d_{cci}d_{aaai} - 2d_{aci}d_{aaci}}{2(d_{aai}d_{cci} - d_{aci}^2)}\, O_p(T^{-1/2}) \\ & - \frac{E\left(\frac{1}{T}\,\mathbf{d}_{\gamma ai0}\right) E\left(\frac{1}{T}\,d_{aci0}\right) - E\left(\frac{1}{T}\,d_{aai0}\right) E\left(\frac{1}{T}\,\mathbf{d}_{\gamma ci0}\right)}{E\left(\frac{1}{T}\,d_{aai0}\right) E\left(\frac{1}{T}\,d_{cci0}\right) - E\left(\frac{1}{T}\,d_{aci0}\right)^2} \\ & \frac{d_{cci}d_{aaci} + d_{aai}d_{ccci} - 2d_{aci}d_{acci}}{2(d_{aai}d_{cci} - d_{aci}^2)} \\ & - \frac{d_{cci}d_{aaci} + d_{aai}d_{ccci} - 2d_{aci}d_{acci}}{2(d_{aai}d_{cci} - d_{aci}^2)}\, O_p(T^{-1/2}) \\ & - \frac{d_{aai}\mathbf{d}_{\gamma cci} + d_{cci}\mathbf{d}_{\gamma aai} - 2d_{aci}\mathbf{d}_{\gamma aci}}{2(d_{aai}d_{cci} - d_{aci}^2)} \end{aligned} \tag{40}$$

Evaluating at γ_0, using the fact that $\eta_i(\gamma) = \eta_{i0} + O_p(T^{-1/2})$, adding $1/T^2$ in numerators and denominators and replacing by expectations:

(41)

$$
\begin{aligned}
B = & -\frac{1}{2}\frac{1}{\left(E\left(\frac{1}{T}d_{aai0}\right)E\left(\frac{1}{T}d_{cci0}\right) - E\left(\frac{1}{T}d_{aci0}\right)^2\right)^2} \\
& \left\{\left[E\left(\frac{1}{T}\mathbf{d}_{\gamma ci0}\right)E\left(\frac{1}{T}d_{aci0}\right) - E\left(\frac{1}{T}d_{cci0}\right)E\left(\frac{1}{T}\mathbf{d}_{\gamma ai0}\right)\right]\right. \\
& \left[E\left(\frac{1}{T}d_{aai0}\right)E\left(\frac{1}{T}d_{acci0}\right) + E\left(\frac{1}{T}d_{cci0}\right)E\left(\frac{1}{T}d_{aaai0}\right) - 2E\left(\frac{1}{T}d_{aci0}\right)E\left(\frac{1}{T}d_{aaci0}\right)\right] \\
& + \left[E\left(\frac{1}{T}\mathbf{d}_{\gamma ai0}\right)E\left(\frac{1}{T}d_{aci0}\right) - E\left(\frac{1}{T}d_{aai0}\right)E\left(\frac{1}{T}\mathbf{d}_{\gamma ci0}\right)\right] \\
& \left.\left[E\left(\frac{1}{T}d_{cci0}\right)E\left(\frac{1}{T}d_{aaci0}\right) + E\left(\frac{1}{T}d_{aai0}\right)E\left(\frac{1}{T}d_{ccci0}\right) - 2E\left(\frac{1}{T}d_{aci0}\right)E\left(\frac{1}{T}d_{acci0}\right)\right]\right\} \\
& - \frac{1}{2}\frac{1}{E\left(\frac{1}{T}d_{aai0}\right)E\left(\frac{1}{T}d_{cci0}\right) - E\left(\frac{1}{T}d_{aci0}\right)^2} \\
& \left[E\left(\frac{1}{T}d_{aai0}\right)E\left(\frac{1}{T}\mathbf{d}_{\gamma cci0}\right) + E\left(\frac{1}{T}d_{cci0}\right)E\left(\frac{1}{T}\mathbf{d}_{\gamma aai0}\right) - 2E\left(\frac{1}{T}d_{aci0}\right)E\left(\frac{1}{T}\mathbf{d}_{\gamma aci0}\right)\right]
\end{aligned}
$$

(42)

$$
+ O_p(T^{-1/2})
$$

Finally, taking the expected value of this expression will not change anything, except that the remainder would be $O(T^{-1})$ instead of $O_p(T^{-1/2})$.

Part C. To analyze C, we need the following result:

$$
\frac{\partial}{\partial a_i}E\left(\mathbf{d}_{\gamma ci}\right) = E\left(\mathbf{d}_{\gamma aci}\right) + E\left(\mathbf{d}_{\gamma ci}d_{ai}\right) \tag{43}
$$

This works with other derivatives of expectations as well.

We are interested in the following derivative, which we will call C^a:

$$
\begin{aligned}
C^a &= -\frac{\partial}{\partial a_i}\left(\frac{E(\mathbf{d}_{\gamma ci})E(d_{aci}) - E(d_{cci})E(\mathbf{d}_{\gamma ai})}{E(d_{aai})E(d_{cci}) - [E(d_{aci})]^2}\right) \\
&= -\frac{\frac{\partial}{\partial a_i}\left(E(\mathbf{d}_{\gamma ci})E(d_{aci}) - E(d_{cci})E(\mathbf{d}_{\gamma ai})\right)}{E(d_{aai})E(d_{cci}) - [E(d_{aci})]^2} \\
&\quad + \frac{\left(E(\mathbf{d}_{\gamma ci})E(d_{aci}) - E(d_{cci})E(\mathbf{d}_{\gamma ai})\right)\frac{\partial}{\partial a_i}\left(E(d_{aai})E(d_{cci}) - [E(d_{aci})]^2\right)}{\left(E(d_{aai})E(d_{cci}) - [E(d_{aci})]^2\right)^2}
\end{aligned}
$$

Working with the derivative and using the above rule, we get:

$$
\begin{aligned}
C^a = & -\frac{1}{E(d_{aai})E(d_{cci}) - [E(d_{aci})]^2} \\
& \{E(\mathbf{d}_{\gamma ci})\left[E(d_{aaci}) + E(d_{aci}d_{ai})\right] + E(d_{aci})\left[E(\mathbf{d}_{\gamma aci}) + E(\mathbf{d}_{\gamma ci}d_{ai})\right] \\
& -E(d_{cci})\left[E(\mathbf{d}_{\gamma aai}) + E(\mathbf{d}_{\gamma ai}d_{ai})\right] - E(\mathbf{d}_{\gamma ai})\left[E(d_{acci}) + E(d_{cci}d_{ai})\right]\} \\
& + \frac{E(\mathbf{d}_{\gamma ci})E(d_{aci}) - E(d_{cci})E(\mathbf{d}_{\gamma ai})}{\left(E(d_{aai})E(d_{cci}) - [E(d_{aci})]^2\right)^2} \\
& \{E(d_{aai})\left[E(d_{acci}) + E(d_{cci}d_{ai})\right] + E(d_{cci})\left[E(d_{aaai}) + E(d_{aai}d_{ai})\right] \\
& -2E(d_{aci})\left[E(d_{aaci}) + E(d_{aci}d_{ai})\right]\}
\end{aligned}
$$

Likewise, for C^c we have:

$$
\begin{aligned}
C^c = & -\frac{1}{E(d_{aai})E(d_{cci}) - [E(d_{aci})]^2} \\
& \{E(\mathbf{d}_{\gamma ai})\left[E(d_{acci}) + E(d_{aci}d_{ci})\right] + E(d_{aci})\left[E(\mathbf{d}_{\gamma aci}) + E(\mathbf{d}_{\gamma ai}d_{ci})\right] \\
& -E(d_{aai})\left[E(\mathbf{d}_{\gamma cci}) + E(\mathbf{d}_{\gamma ci}d_{ci})\right] - E(\mathbf{d}_{\gamma ci})\left[E(d_{aaci}) + E(d_{aai}d_{ci})\right]\} \\
& + \frac{E(\mathbf{d}_{\gamma ai})E(d_{aci}) - E(d_{aai})E(\mathbf{d}_{\gamma ci})}{\left(E(d_{aai})E(d_{cci}) - [E(d_{aci})]^2\right)^2} \\
& \{E(d_{cci})\left[E(d_{aaci}) + E(d_{aai}d_{ci})\right] + E(d_{aai})\left[E(d_{ccci}) + E(d_{cci}d_{ci})\right] \\
& -2E(d_{aci})\left[E(d_{acci}) + E(d_{aci}d_{ci})\right]\}
\end{aligned}
$$

We then evaluate at γ_0 and take the expected value of these expressions. Putting everything together. If we, finally, add all the terms of B and C from before, which is equal to $\mathbf{d}_{\gamma Mi}(\gamma) - \mathbf{d}_{\gamma i}(\gamma, \eta_i(\gamma)) = B + C$, we get exactly minus (32). Therefore, the modified score equal the standard score minus the first order term of the bias, because we are substracting it with the modification $B + C$. The reminder of this expansion for $\mathbf{d}_{\gamma Mi}(\gamma)$ is $O(T^{-1})$, as opposed to $O(1)$ that is the order of magnitude of the bias of $\mathbf{d}_{\gamma i}(\gamma, \eta_i(\gamma))$. This shows that MMLE reduced the order of the bias of the MLE.

Bibliography

Arellano, M., and J. Hahn (2006): "A likelihood-based approximate solution to the incidental parameter problem in dynamic nonlinear models with multiple effects," Unpublished Manuscript.

——— (2007): "Understanding Bias in Nonlinear Panel Models: Some Recent Developments," in *Advances in Economics and Econometrics, Theory and Applications, Ninth World Congress*, ed. by R. Blundell, W. Newey, and T. Persson, pp. 105–142. Cambridge University Press, New York.

Bester, C. A., and C. Hansen (2009): "A Penalty Function Approach to Bias Reduction in Non-linear Panel Models with Fixed Effects," *Journal of Business and Economic Statistics*, 27(2), 131–148.

Carro, J. M. (2007): "Estimating dynamic panel data discrete choice models with fixed effects," *Journal of Econometrics*, 140(2), 503–528.

Carro, J. M., and A. Traferri (2009): "Correcting the bias in the estimation of a dynamic ordered probit with fixed effects of self-assessed health status," *Economics Working Paper series. Universidad Carlos III de Madrid*, (09-40).

Chamberlain, G. (1984): "Panel Data," in *Handbook of Econometrics*, ed. by Z. Griliches, and M. Intriligator, vol. 2. Elsevier Science, Amsterdam.

Contoyannis, P., A. M. Jones, and N. Rice (2004): "The Dynamics of Health in the British Household Panel Survey," *Journal of Applied Econometrics*, 19(4), 473–503.

Fernandez-Val, I. (2009): "Fixed effects estimation of structural parameters and marginal effects in panel probit models," *Journal of Econometrics*, 150(1), 71–85.

Greene, W., and D. Hensher (2010): *Modeling ordered choices: A primer*. Cambridge University Press, Cambridge, UK.

Hahn, J., and G. Kuersteiner (2004): "Bias Reduction for Dynamic Nonlinear Panel Models with Fixed Effects," Unpublished Manuscript.

Hahn, J., and W. Newey (2004): "Jackknife and Analytical Bias Reduction for Nonlinear Panel Models," *Econometrica*, 72(4), 1295–1319.

Jones, A. M. (2007): "Panel Data Methods and Applications to Health Economics," in *The Palgrave Handbook of Econometrics Volume II: Applied Econometrics*, ed. by T. C. Mills, and K. Patterson, vol. 2. Palgrave MacMillan, Basingstoke, UK.

Lindeboom, M., and E. Van Doorslaer (2004): "Cut-point shift and index shift in self-reported health," *Journal of Health Economics*, 23(6), 1083–1099.

Mundlak, Y. (1978): "On the pooling of time series and cross section data," *Econometrica*, 46(1), 69–85.

van Doorslaer, E., X. Koolman, and A. Jones (2004): "Explaining income-related inequalities in doctor utilisation in Europe," *Health economics*, 13(7), 629–647.

Wooldridge, J. (2005): "Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity," *Journal of Applied Econometrics*, 20(1), 39–54.

CHAPTER 3

Gender differences in Major Choice and College Entrance Probabilities in Brazil

Abstract. I study gender differences in major choice and college entrance probabilities in University of Campinas, a Brazilian public university dependent on the State of São Paulo. As with most Brazilian public universities, students select a major, and then compete for a place in that major by taking a major-specific entrance exam. This singular characteristic of the Brazilian case allows me to differentiate the effect of gender on major-specific entrance probabilities and preferences. I propose a model and econometric strategy which can account for two important issues, selectivity bias and the fact that expected utility depends on the probability of entering the different majors. I find evidence of gender differences in preferences and entrance probabilities. For most majors, gender differences in major choice are mostly explained by differences in preferences. However, for the most demanding majors (those that require higher grades from students), differences in major choice are explained in a large proportion by differences in entrance probabilities. Finally, I find that gender has important interactions with other variables. In particular, gender effects depend on education, socioeconomic characteristics and family background.

1. Introduction

There are significant differences in the average major choices of men and women. This issue has been analyzed extensively for the American case (see, for example Freeman 1971, Turner and Bowen 1999, Zafar 2009), but the same pattern is present in Brazilian universities. In the case of the University of Campinas, for example, the proportion of men choosing engineering majors between 2006 and 2008 was 67 percentage points higher than the proportion of women choosing that major, and the proportion of women choosing Medicine was 28 percentage points higher than the proportion of men choosing that major (see Table 3 for more details).

The choice of college major depends not only on utility considerations (including expected earnings), but also on the relative advantages of each student. *Ceteris paribus*, students with a relative advantage in math will tend to do better in majors which make intensive use of mathematics, and therefore we expect them to choose engineering majors more often than students who have a relative advantage in verbal skills, for example.

Accordingly, men and women may choose different majors partly because they have different preferences or different relative advantages[1]. Moreover, gender differences may be affected by education, socioeconomic characteristics and family background. For example, gender differences in major choice may be smaller or larger for students who attended public schools, in comparison with students who attended private schools.

Therefore, it is important to know what is the relative importance of gender differences and other individual characteristics in explaining major choice and academic performance. Does gender have an effect on the probability of entering college once we take into account other individual characteristics? Do men and women choose different majors because they have different preferences or different relative advantages, or because they have received a different education and come from different socioeconomic backgrounds?

In past works, it has been difficult to find an answer to the previous questions because in most countries students are allowed to enter college (or not) before choosing major. Therefore, it is in general not possible to discern whether the major choice was motivated by the student preferring that choice more than others, or because the student believes she will do better in that major.

A good survey of previous works is Turner and Bowen (1999). Their discussion on the possible causes of gender differences makes clear that it is generally very difficult to determine the differential impact of each factor. They use data on SAT

[1]Several hypothesis have been proposed in the literature as to why these gender differences may arise. The fertility hypothesis says that women know that their work life will be interrupted when they have children, so the economic value of particular careers may be lower. The socialization hypothesis states that men and women are taught to be different since their infancy, and this has an effect on what they perceive their role in society should be. Finally, gender discrimination may also play a role, not only by generating a difference in expected earnings for particular careers, but also by pre-empting the entry of women (men) to careers traditionally dominated by men (women). See Turner and Bowen 1999 for a more detailed discussion. However, it is important to remark that, independently of the reasons why gender differences in preferences or academic performance, it is important to know whether these differences do, in fact, exist.

(Scholastic Aptitude Test) scores for American college students, and show that an important part of the gender gap is explained by differences in SAT scores. Other papers, like Altonji (1993) and Arcidiacono (2004), present dynamic models to study the choice of college and major. However, they are mainly concerned with the effects of differences in predicted earnings, and do not take into account differences in the probability of entering or finishing the different majors.

An important precedent for this paper is Montmarquette et al. (2002), who use an econometric approach which is closer to the one used in this paper. They present a model in which students take into account the probability of graduating when choosing major. They estimate this probability through a linear probability model, and then introduce the estimated probability in a multinomial logit of choice of major. However, they do not allow for correlation between the errors of the probability equation and the expected utility of each major, which implies that their analysis is prone to sample selection bias. As I will show, an alternative econometric strategy can account for correlation between both equations.

In this sense, the Brazilian case is particularly interesting. In most Brazilian public universities, the student chooses a major before taking a major-specific exam, which determines whether the student is allowed to enroll in the major or not. Therefore, when choosing among the different majors, the student takes into account not only the utility corresponding to each major, but also the associated probabilities of entry. This implies that we can perform a separate study of the factors affecting the probability of entry and the choice of major.

The determinants of major choice have been scarcely explored in the Brazilian case. A few papers analyze the determinants of performance in Entrance Test Exams (Guimarães and Sampaio 2007, 2008, Calvacanti et al. 2009), but the choice of major has not been analyzed in detail. Such analysis of gender differences in choice of major is important because of its possible relation with gender inequality. This is an important topic of research for the Brazilian Federal Government, which is currently designing public policies to reduce gender inequality. For example, the Federal Government has recently introduced over 400 projects directed at enhancing equal opportunities for men and women, which will be performed by 22 government institutions between 2008 and 2011 (Pinheiro et al. 2008).

In this paper, I estimate a model of major choice and college entry using data from entrance tests of the University of Campinas between 2006 and 2008. The

basic model (Model I) consists of two estimations. First, I estimate a binary logit to study the determinants of the probability of entering the different majors. Then, I estimate a multinomial logit model of major choice. An important determinant of expected utility is the probability of entering the major. I calculate these probabilities from the estimations of the first step.

In the basic model, I assume the errors of the entry and expected utility equations are uncorrelated, so there is no selectivity bias by assumption. In other words, the fact that a student chooses a particular major does not mean that this student has a higher probability of entering that major, in comparison with a similar student who chose another major. In an extension (Model II), I allow for correlations between the errors of both equations, so that students with a larger entry shock for a particular major tend to have a larger preference shock for the same major. Model I is estimated through Maximum Likelihood in two steps. Model II is estimated through Maximum Simulated Likelihood in one unique step.

When estimating the second model, I find that the correlation between the errors of the two equations is positive and significant. Therefore, students who get a larger preference shock for some major tend to have a larger entry shock for that major too. Given that the coefficient is significant, the model without correlations will produce biased estimators. Therefore, it is important to consider correlated errors in the econometric design.

I find evidence of gender differences in the probability of entering the different majors. Controlling for other individual characteristics, men have on average a higher probability of entering some majors, and women have a higher probability of entering other majors. Interestingly, the effect of gender depends on past academic performance, given that for most majors, the interaction between gender and the ENEM grade is significant.[2]

I also find a significant effect of gender on major choice. The largest differences between men and women arise in the most demanding majors (i.e. those with highest minimum required grades). Nevertheless, there are significant differences between the choices of men and women in other majors as well. Overall, men have a higher probability of choosing mathematically-oriented majors (Technologies,

[2]ENEM (Exame Nacional do Ensino Médio, High School National Exam) is a non-mandatory Brazilian national exam, which examines students' knowledge of concepts taught in secondary school.

Exact Sciences, and Engineering and Architecture), and women have a higher probability of choosing majors in Natural and Earth Sciences, Arts, Humanities, and Health and Biological Sciences. In addition, I find that the effect of gender on major choice depends on student characteristics. In particular, the size of gender differences depends on work status, type of secondary school and income, among other variables.

In order to determine if gender differences in major choice are caused by differences in preferences or probability of entry, I simulate women choices with male probabilities of entry, and men choices with female probabilities of entry. I find that preferences account for most of the difference in choices in majors with low or medium minimum required grades. In the most demanding majors, on the other hand, a large part of the difference in major choice is explained by differences in the probability of entry.

The rest of the paper proceeds as follows. In Section 2, I present the models to be estimated and the estimation strategy. In Section 3, I describe the process of major selection and the characteristics of the entrance exam for University of Campinas. In Section 4, I present an introduction to the data and show descriptive statistics of the sample and variables used in the estimations. In Section 5, I show and discuss the results of the estimations. Finally, Section 6 presents the conclusions of the paper.

2. Model and estimation strategy

In this section, I present a model to study the decisions of N individuals ($i = 1, 2, \dots N$), choosing among J majors ($j = 1, 2, \dots J$). Let $M_{ij} \in \{0, 1\}$ be a binary indicator, which is equal to 1 if individual i chooses major j and 0 otherwise, and $E_{ij} \in \{0, 1\}$ be another binary indicator, which is 1 if individual i enters major j and 0 otherwise. Clearly, M_{ij} will be equal to 1 for exactly one major j, and we observe E_{ij} only for the major the individual has actually chosen.

Individuals choose a major in order to maximize their expected indirect utility, which depends on the utility of entering the major and the probability of entering the major, which in turn depend on individual characteristics. Specifically, let $U_{ij} = \alpha_j x_i$ be the utility of individual i of entering major j, and let

$p_{ij} = Pr(E_{ij} = 1|x_i)$ be the probability of entering major j, where x_i is a k_x-vector of individual characteristics (including a 1 for the intercept), and the α_j are k_x-vectors of parameters.

The expected indirect utility of individual i from choosing major j is:

$$u^*_{ij} = p_{ij}\, U_{ij} + \varepsilon_{ij}, \tag{44}$$

where ε_{ij} is an individual-major taste shock, which is unobserved by the econometrician, but known to the individual when choosing among the different majors. Introducing the expression for utility and rearranging equation (44) we get:

$$u^*_{ij} = p_{ij}\, \alpha_j\, x_i + \varepsilon_{ij}. \tag{45}$$

According to equation 44, the utility assigned to a given major depends on the probability of entering that major. This probability depends on the characteristics of the individual, and is determined by a binary model with latent equation:

$$y^*_{ij} = \gamma_j\, z_i + \eta_{ij}, \tag{46}$$

where z_i is a k_z-vector of individual characteristics, possibly overlapping with x_i, γ_j is a vector of major-specific parameters, and η_{ij} is an error term which is unobserved (or partially observed) by the individual when choosing major, and unobserved by the econometrician before and after the individual chooses major. z_i includes a 1 for the intercept.

As usual, we cannot observe the latent variables u^*_{ij} and y^*_{ij}. The rules determining the observed variables are:

$$\begin{aligned} M_{ij} &= 1\left[u^*_{ij} > \max_{k=1,\ldots J,\, k\neq j} u^*_{ik}\right], \\ E_{ij} &= 1\left[y^*_{ij} > 0\right], \end{aligned}$$

where $1[\,.\,]$ is an indicator function.

There are two difficulties in estimating the above model. First, we only observe E_{ij} when $M_{ij} = 1$. Therefore, there will be a selection process if the errors in equations (44) and (46) are correlated. Second, the first latent equation depends on the parameters of the second latent equation, through the probability of entering the major.

2.1. Basic model. To complete the description of the model, we need some assumption about the error terms. I will start by assuming independence of the error terms (Model I), and then relax this assumption by introducing correlations between the error terms (Model II). The benchmark model is characterized by the following assumption:

ASSUMPTION 1 (Model I). *ε_{ij} are i.i.d. according to a double exponential distribution, and have zero mean and variance equal to $\pi^2/6$. η_{ij} are i.i.d. with a cumulative density function F, and have zero mean and unit variance. ε_{ij} and η_{ik} are independent for any j and k.*

The first part of Assumption 1 corresponds to what is known as the multinomial logit Model (MNL, McFadden 1974).[3] Alternatively, I could have assumed a multivariate normal distribution for ε_{ij}, which would have yielded a multinomial probit Model (MNP). The advantage of the MNL is that it provides closed form solutions for the probabilities, and is therefore more tractable (the MNP usually requires numerical integration for solving multiple integrals, which becomes unfeasible when the number of alternatives is large). As is well known, the main disadvantage of the MNL is the IIA (Independence of Irrelevant Alternatives) property.[4]

I will use the MNL for two reasons. First, it allows for comparison with previous studies of major choice. Turner and Bowen (1999), Montmarquette et al. (2002) and Arcidiacono (2004), for example, use the MNL as their discrete choice model. Second, in the following section, I will generalize the model to allow correlations between the errors of equations (44) and (46), which will eliminate the IIA property. Using the MNL for this baseline estimation allows me to compare its results with the alternative specification.

[3]Some authors also refer to this model as the Conditional logit Model, but it is more appropriate to use the term multinomial logit for the case in which the model is derived from utility maximization.

[4]The IIA property requires that the relative odds ratio between two alternatives does not change when a new alternative is added to the set of alternatives or when the characteristics of a third alternative change. In the case of the MNL, the ratio of probabilities of two events is

$$\frac{P_{ij}}{P_{ik}} = \frac{\exp(p_{ij}\,\alpha_j\,x_i)}{\exp(p_{ik}\,\alpha_k\,x_i)}.$$

It is easy to see that this ratio does not depend on the utility parameters of the other choices, which implies that the MNL has the IIA property. See Ben-Akiva and Lerman (1985) and Anderson, De Palma, and Thisse (1992) for more details.

With respect to the distribution of η_{ij}, I could use a normal (probit) or logistic (logit) distribution. The choice between the binary logit and probit models is largely one of convenience and convention, since the substantive results are generally indistinguishable. For the purpose of this paper, I will use the logistic distribution, because of its tractability.

Model I is the easiest model that can be estimated. The errors are independent, which means there is no selection problem. In other words,

$$Pr(E_{ij} = 1|x_i, M_{ij} = 1) = Pr(E_{ij} = 1|x_i),$$

and the same is true for the expectations. It is important to understand the meaning of this assumption: the fact that an individual chooses a given major does not give any information as to whether he has a higher probability of entering that major than other similar students who are not choosing that major.

Under Assumption 1, the probability of entering major j is

$$\begin{aligned} p_{ij} &= Pr(E_{ij} = 1|x_i) \\ &= Pr(\gamma_j z_i + \eta_{ij} > 0) \\ &= \frac{1}{1 + \exp(-\gamma_j z_i)}, \end{aligned} \tag{47}$$

and the probability of choosing major j is

$$\begin{aligned} P_{ij} &= Pr(M_{ij} = 1|x_i) \\ &= \frac{\exp(p_{ij}\,\alpha_j\,x_i)}{\sum_{j=1}^{J}\exp(p_{ij}\,\alpha_j\,x_i)}. \end{aligned} \tag{48}$$

The estimation approach is very simple. Given that the parameters of the probability of entering a major are needed to determine the probability of choosing that major, I first run a binary logit of E_{ij} on z_i, for each major j using only the observations of individuals who choose that specific major. Then, we use these estimations to run a multinomial logit of M_{ij} on $p_{ij}\,x_i$, using all observations.

It is important to remark that the proposed two-step procedure will give unbiased estimators only if Assumption 1 is correct. If the errors of the choice and entry equations are correlated, then there will be a selection process which the two-step procedure will not take into account, producing biased estimators. Notice, however, that this has been the approach taken by previous papers examining gender differences in major choice. In the next section, I present an alternative

approach which will yield unbiased estimators for a specific correlation structure. Designing an econometric model and estimation strategy to account for more general correlation structures is an interesting topic for further research, but is beyond the scope of the present paper.

In terms of identification, the parameters of the model are already identified because of the non-linear functional form of probability, and because entrance probabilities enter utility in a multiplicative form. However, as I describe in Section 4 I will include exclusion variables in the choice and entry equations, which will also help in the identification of the parameters. In particular, some variables, like the one describing whether the student has taken a private preparation course for the entrance test, affect only the probability of entering a major. Other variables, like the variables describing the reasons why the student chose the particular major, affect only preferences.

2.2. Model with correlations. In this section we study what happens if ε_{ij} and η_{ij} are correlated. Specifically, let $\varepsilon_{ij} = \nu_{ij} + \sigma\,\mu_{ij}$ and $\eta_{ij} = e_{ij} + \sigma\,\mu_{ij}$, where μ_{ij} is the common factor affecting the taste and entry shocks. If σ is significant, then students with a higher entry shock for some major tend to have a higher taste shock for that major too.

It is important to determine what the individual and the econometrician know before choosing a major. ν_{ij} and μ_{ij} are known by the individual before choosing a major, but are unobserved by the econometrician. e_{ij} is not known by the individual (at least before taking the exam), nor by the econometrician (before and after the exam).[5]

ASSUMPTION 2 (Model II). *ν_{ij} are i.i.d. according to a double exponential distribution, with zero mean and variance $\pi^2/6$. e_{ij} are i.i.d. with cumulative density function F, and have zero mean and unit variance. μ_{ij} are i.i.d. according to a cumulative density function G with mean 0, probability density function g, and unit variance. μ_{ij}, ν_{ik} and e_{ih} are independent for any j, k, and h.*

The econometric approach is inspired in the mixed logit, which allows the parameters to differ among individuals.[6] McFadden and Train (2000) show that

[5]An alternative way to introduce correlation would have been to assume that the ε_{ij}'s and η_{ij}'s have a joint normal distribution, and to allow the covariance matrix to be non-diagonal.
[6]Read chapter 6 of Train (2003) for a good description of the mixed logit.

the mixed logit can approximate any discrete choice process, by appropriately choosing the distribution G, and that the mixed logit eliminates the well known problem of the Independence of Irrelevant Alternatives (IIA) of the Conditional and multinomial logit models.[7] Usually, a normal distribution is used when the parameters can take positive or negative values, and a lognormal distribution is used when the parameters must have a specific sign. For the purposes of this paper, μ_{ij} will follow a normal distribution.

Estimation of Model II is not as straightforward as the previous case. The problem is that the correlation between the errors implies that $Pr(E_{ij} = 1|x_i, M_{ij} = 1) \neq Pr(E_{ij} = 1|x_i)$, so the estimation of the second equation using only the observations of individuals who chose a particular major gives biased estimates. For this reason, the two-step procedure cannot be used and we have to estimate the whole system in one step.

The first step is to construct the likelihood function. Let μ_i be a J-vector containing the common factors for all majors, and suppose μ_i is observed by the econometrician. This means that μ_i becomes a variable in the estimation, just as one of the x_i or z_i. Let $P_{ij}(\mu_i) = Pr(M_{ij} = 1|x_i, \mu_i)$ and $p_{ij}(\mu_{ij}) = Pr(E_{ij} = 1|x_i, \mu_{ij})$.[8] Under our assumptions, we have that:

$$Pr(E_{ij} = 1|x_i, \mu_{ij}) = Pr(E_{ij} = 1|x_i, M_{ij} = 1, \mu_{ij}). \tag{49}$$

According to equation 49, if the model is correctly specified (i.e. if the errors of the choice and entry equations are correlated according to Assumption 2) and we were able to observe μ_i, then there would not be a selection problem when estimating the probabilities of entry. In other words, if Assumption 2 is correct, once we control for μ_i, the fact that a student chooses a major does not mean that she has a different probability of entering that major than a student with identical x_i, z_i and μ_i who did not chose that major.

[7] Under Assumption 2, the ratio of probabilities of two events is

$$\frac{\mathcal{P}_{ij}}{\mathcal{P}_{ik}} = \frac{\int_{-\infty}^{\infty} \frac{\exp(p_{ij}\,\alpha_j\,x_i)}{\sum_{j=1}^{J}\exp(p_{ij}\,\alpha_j\,x_i)}\,g(\mu_i)\,d\mu_i}{\int_{-\infty}^{\infty} \frac{\exp(p_{ik}\,\alpha_k\,x_i)}{\sum_{j=1}^{J}\exp(p_{ij}\,\alpha_j\,x_i)}\,g(\mu_i)\,d\mu_i},$$

where μ_i is a J-vector containing the $\mu_i j$ for $j = 1, \ldots J$. Clearly, this ratio depends on the utility parameters of the other choices, which implies that the IIA property does not hold under Assumption 2.

[8] Notice that P_{ij} depends on the common factors for all majors (μ_i), while p_{ij} depends only on the shock for major j (μ_{ij}).

There are $2J$ possible events for which we have to find a probability: the probability of choosing major j and entering, and the probability of choosing major j and not entering, for each j. Given μ_i, the probability of choosing major j and the probability of entering major j are

$$\begin{aligned} P_{ij}(\mu_i) &= \frac{\exp(p_{ij}\,\hat{\alpha}_j\,x_i + \sigma\,\mu_{ij})}{\sum_{j=1}^{J}\exp(p_{ij}\,\hat{\alpha}_j\,x_i + \sigma\,\mu_{ij})} \\ p_{ij}(\mu_{ij}) &= F(\gamma_j\,z_i + a_i + \sigma\,\mu_{ij}). \end{aligned}$$

Then, the probability of choosing major j and entering is $P_{ij}(\mu_{ij})\,p_{ij}(\mu_{ij})$, and the probability of choosing major j and not entering is $P_{ij}(\mu_{ij})\,(1-p_{ij}(\mu_{ij}))$.

Of course, we do not observe μ_i. Nevertheless, we know its distribution, so we can integrate out the μ_i, i.e. calculate the expected value of the above probabilities:

$$LE_{ij} = \int_{-\infty}^{\infty} P_{ij}(\mu_i)\,p_{ij}(\mu_{ij})\,g(\mu_i)\,d\mu_i \tag{50}$$

$$LN_{ij} = \int_{-\infty}^{\infty} P_{ij}(\mu_i)\,(1-p_{ij}(\mu_{ij}))\,g(\mu_i)\,d\mu_i. \tag{51}$$

With LE_{ij} and LN_{ij} we obtain the following log-likelihood function:

$$L = \sum_{i=1}^{N}\sum_{j=1}^{J} M_{ij}\;(E_{ij}\;\log(LE_{ij}) + (1-E_{ij})\;\log(LN_{ij}))$$

There is no closed-form solution for the above integrals, and numerical integration is unfeasible when the number of majors is large.[9] Therefore, I will approximate the above probability through simulations and maximize the simulated log-likelihood function.

The estimation process is as follows. For a given value of the parameters, a realization of μ_i is drawn from G for each individual. Using these draws, I calculate P_{ij} and p_{ij}. The process is repeated for R draws and yields the following

[9]In this paper, for example, I will group the majors in 9 major concentrations, which would require solving 9 integrals for each of the above probabilities.

approximate probabilities:

$$\breve{LE}_{ij} = \frac{1}{R}\sum_{r=1}^{R} \breve{P}_{ij}^{r}\,\breve{p}_{ij}^{r}$$

$$\breve{LN}_{ij} = \frac{1}{R}\sum_{r=1}^{R} \breve{P}_{ij}^{r}\,(1-\breve{p}_{ij}^{r}),$$

where $\breve{P}_{ij}^{r}$ and $\breve{p}_{ij}^{r}$ are the simulated probabilities of choosing and entering major j corresponding to draw r.

Using these simulated probabilities we obtain the following simulated log-likelihood function:

$$\breve{L} = \sum_{i=1}^{N}\sum_{j=1}^{J} M_{ij}\,\left(E_{ij}\,\log(\breve{LE}_{ij}) + (1-E_{ij})\,\log(\breve{LN}_{ij})\right).$$

The Maximum Simulated Likelihood (MSL) estimator simply maximizes the above simulated log-likelihood, and is obtained through an iterative maximization algorithm as the usual ML estimator. The only difference is that in each step, we use a particular draw of the random term μ_{ij} to simulate the probabilities in order to construct the objective function. With respect to the choice of R, the estimators will be asymptotically consistent if R grows at a rate greater or equal than $\sqrt{N}$. Therefore, in applications R is usually chosen to be slightly larger than $\sqrt{N}$.

2.3. Alternative models. In this section, I discuss alternative models which could have been used to analyze major choice.

First, as I mentioned earlier, an alternative to the multinomial logit (MNL) is the multinomial probit (MNP). As is well known, the MNP does not have the problem of the IIA property. On the other hand, the MNL has the advantage of delivering closed form solutions for probabilities, which reduces the computational burden of the estimations. This property is very important for the present paper because the number of alternatives and observations is very large. Moreover, the model with correlations (Model II) eliminates the IIA property of the standard MNL, so the main argument to use the MNP instead of the MNL loses strength.

Second, another econometric approach which could have been used is the one proposed by Mallar (1977). Mallar studied a model with a set of interrelated dichotomous (binary) relationships, where the probability that one event happens

affects the probability that other events occur. The approach is to transform the model, so that each probability is a non-linear function of a linear index. Then, assuming that each linear index depends on the other linear indexes (rather than on the other probabilities), it is possible to obtain a reduced form for the linear indexes, and estimate an independent equation for each probability, in which the dichotomous variable depends only on the exogenous variables (i.e. the model is transformed so that each probability does no longer depend on the other probabilities). Mallar shows that the structural parameters can be obtained from the parameters of the reduced form.

We could interpret a polychotomous (multinomial) model as a model in which the n choices are interrelated dichotomous events. Then, we could apply Mallar's approach to our problem, by adding the respective probabilities of entry. However, the above referenced transformation is a strong assumption for a multinomial model. Mallar's approach is specially suitable for the case where *truly* binary variables depend on one another, but is less satisfying for the analysis of polychotomous choices, for which the MNL and MNP have been specially designed.

In addition to the previous reasons, an important advantage of the MNL for the analysis of major choice is that this is the model which has been used the most in the literature, so it facilitates comparison with other papers. For these reasons, I will use the econometric model proposed in the previous sections to perform the analysis.

3. The university entrance process

Before describing the data, it is important to understand the entrance process for Brazilian universities (vestibular). In this section, I describe the process for University of Campinas, but the process is similar for most public universities. I will only present a brief description of the process of choosing a major and entering the university, considering the most relevant features for the present paper. The actual entrance process is much more complex.[10] Basically, university candidates must choose their preferred majors, and only the best students within each major are called to fill the seats. Majors with more candidates per offered seat are more competitive and thus imply a lower probability of entry.

[10]For more details, read University of Campinas' general resolutions (resoluções gerales) number 31 of August 10, 2005; number 41 of August 21, 2006; and number 30 of August 8, 2007.

Candidates must follow the following process:

(1) Each candidate chooses 3 majors in order of preference.

(2) Candidates take the first-stage exam. The exam is the same for all students and has 2 parts: multiple choice questions and essay. The multiple choice questions may belong to Mathematics, Physics, Chemistry, Biology, History or Geography. The essay evaluates knowledge of the Portuguese language, and is graded only for students who answered correctly at least 50% of the multiple choice questions. All students are ordered within their major of first-choice. Only the top students within each major can participate in the second-stage of the exam.[11] It is important to remark that in this first-stage of the exam students compete only with other students *choosing the same major* for the possibility of taking the second-stage exam.

(3) Candidates take the second-stage exam. The exam is the same for all students, and has 8 parts: Literature and Portuguese Language, Biological Sciences, Geography, History, Mathematics, Chemistry, Physics, and English language.[12] The different parts of the exam are given different weights for each major (e.g. engineering majors put more weight on mathematics, literature puts more weight on Portuguese language).

(4) Students with the highest scores are called to fill a seat. If the student is not called for her first choice, she may be called for her second or third choice, if her score is high enough. As students may decide not to enroll in a major when they are called, there may be several calls until all seats at the different majors are filled.

(5) Candidates decide whether to enroll or not. Even when a candidate is called to fill a seat, she may decide not to enroll in the university. There are many reasons why this may happen:

   (a) The candidate may be one of the so called "treineiros," students who have not finished high school, and are only training for the exam in a subsequent year.

[11]The number of students going to the second stage is determined following a series of complex rules. The basic intention is to have a maximum of eight candidates per seat in the second-stage exam.

[12]Some majors require an additional aptitude test.

(b) The candidate is called to fill a seat at another university, and chooses that university over this one.

(c) The candidate decides not to enroll for other reasons (e.g. a change in her socioeconomic situation).

In stage (4), students may be offered a place at their second or third choices in an early call, but regardless of whether they accept or reject this offer, they may still be offered a superior choice in later calls. However, students who reject an offer to fill a seat in one of their chosen majors in any given call, will not be offered lower choices in subsequent calls. For example, if in the first call a student is offered a seat at her second choice major, she may reject it and still be offered a seat at her first choice in later calls, but she will no longer be offered a seat at her third choice.

With respect to the enrollment decision, in the available database it is possible to identify treineiros, but it is not possible to identify students deciding not to enroll for other reasons. Treineiros will be excluded from the analysis, because they may have different motives for choosing a major than non-treineiros.[13]

As explained above, students submit an ordered list of 3 majors with their university applications. For the purposes of this paper, I will focus on the study of the determinants of the first choice. There are several reasons for this decision. First, the available database does not have information on students' second and third choices. Second, the proportion of students enrolling in their second or third choice is much smaller than the proportion of students enrolling in their first choice.[14] Table 1 shows the number of candidates, offers to fill seats (calls), and enrollments, depending on the order of the major in the list of preferred majors. The table shows that 90.95% of enrolled students enroll in the major which they

[13]Treineiros want to take the exam in order to gain experience in the vestibular process. Given that the exam is basically the same for all majors (all that changes from major to major is the weight given to each part of the exam), treineiros care less about which major they choose as their first choice. Therefore, many treineiros choose easier majors in order to have a higher probability of getting to the second stage of the exam. In the database, for example, it is possible to see that treineiros choose Technology majors (the least demanding group of majors) in a much higher proportion than non-treineiros.

[14]Even though the database has no information on the majors that students select as second or third alternatives, we know in which majors students are offered a seat, and also in which majors students decide to enroll (if any). Therefore, we can determine if the student receives an offer or enrolls in a major which was her first choice or not. Notice however, that we cannot determine the second or third choice of students who are not offered any seat.

selected as their first choice, and only 9.05% of enrolled students enroll in a major which was not their first choice. Moreover, the share of calls for second and third choices is 14.05%, which means that students reject second and third choices in a higher proportion than first choices. Third, the reasons for choosing the second and third alternatives may be very different from the reasons for choosing the first alternative. For example, the second choice may be a 'safe bet,' that is, a major for which the probability of entry is much higher than the first choice, and which the student selects in order to maximize the probability of entering in some major.

Finally, students may be applying to other universities in addition to University of Campinas. Then, it could be the case that these students select a major at University of Campinas as a 'safe bet,' and choose a different major in the other university. The information on applications to other universities is not available in the database. However, it is unlikely that students will choose a major at University of Campinas as a safe bet, given that University of Campinas is one of the most prestigious universities in Brazil, and it is generally considered to be very difficult to enter this university. For example, over 60% of the students in the sample stated that they chose this university for its reputation, or because this university is the best for the major they want to study.

[TABLE 1 ABOUT HERE.]

4. Data and descriptive statistics

The dataset is composed of major choices, entrance test outcomes and individual characteristics of applicants to the University of Campinas between 2006 and 2008. The database has 134,563 observations (not including treineiros), each corresponding to one candidate. After eliminating observations with missing values, we are left with 120,058 observations.

Table 2 shows the number of candidates for a seat at the University, the number of students called to fill a seat in their first choice major, and the number of students enrolled in their first choice major; separated by gender, where M stands for male and F for female.[15] Interestingly, we can see that the difference in the probability of being offered a seat between men and women is between 2 and 4

[15]All tables in the paper are based on the sample used in the estimations, i.e. they do not include treineiros and students with missing information for some variable used in the estimations. Tables

percentage points, depending on the year. Considering the whole sample, male students' average probability of being called to fill a seat is 12.34%, while female student's probability is 9.57%.

[TABLE 2 ABOUT HERE.]

Majors have been grouped in 9 areas. Groups have been constructed taking into account similarity of the fields of study and degree of difficulty. The degree of difficulty is determined by the test score of the last person called to fill a seat: more difficult majors will require higher grades from students to be called to fill a seat. The major composition of the different groups is the following:

Technologies: Construction Technology, Environmental Sanitation Technology, Telecommunications Technology, Information Technology.

Exact Sciences: Statistics, Mathematics (teaching certificate), Computer Science, Physics (teaching certificate), Physics-Mathematics-Applied Mathematics, Computation.

Engineering and Architecture: Electrical Engineering, Civil Engineering, Chemical Engineering, Mechanical Engineering, Electrical Engineering, Computer Engineering, Control and Automation Engineering, Architecture and Urban Planning.

Natural and Earth Sciences: Geography, Geology-Geography, Chemical Technology, Chemistry, Food Engineering, Agricultural Engineering.

Arts: Music Conducting, Music Composition, Music (teaching certificate), Music Instruments, Popular Music, Dance, Visual Arts, Scenic Arts.

Social Sciences: Social Sciences, Literature, History, Economics, Social Communication (Media Studies).

Humanities: Pedagogy (teaching certificate), Chemistry-Physics (teaching certificate), Linguistics, Language Studies, Language Studies (teaching certificate), Philosophy.

Health and Biological Sciences: Pharmacy, Medicine, Biological Sciences.

Other Health and Biological Sciences: Nursing, Physical Education, Phonology, Dentistry, Biological Sciences (teaching certificate).

---

including treineiros or students with missing information may be obtained from the author upon request.

University of Campinas organizes majors in 4 areas, according to similarity of fields of study: (i) Exact, Technological and Earth Sciences; (ii) Humanities; (iii) Arts; and (iv) Biological and Health Sciences. The approach taken to construct major concentrations was to divide these areas in groups according to the grade of the last person called to fill a seat. Humanities, and Biological and Health Sciences were divided in two groups each. Arts was kept as one group because there was not much heterogeneity between majors in terms of minimum grades. Exact, Technological and Earth Sciences was divided in 4 groups because this was the area with the largest number of majors, and with more heterogeneity in fields of study and minimum grades. Architecture was placed in the same category as most Engineering majors because it belongs to the same faculty as Civil Engineering, and has a similar minimum grade. Food Engineering and Agricultural Engineering were placed separately from the other Engineering majors because they have much lower minimum grades. Instead, these majors were placed in Natural and Earth Sciences, which is composed of majors with similar degree of difficulty and field of study.

The major offering of University of Campinas remained unchanged from 2006 to 2008. University of Campinas gives two kinds of academic degrees. A Bachelor's degree corresponds to a BSc or BA degree in American Universities. A Teaching Certificate is an inferior degree, which usually requires a lower grade to enter and is intended for graduates who want to teach at the secondary level of education (high school). I have indicated which majors correspond to Teaching Certificates in the list of majors. All other majors correspond to Bachelor's degrees.

One difference between this paper and previous works is that I will have to include more groups in the analysis.[16] This difference arises because previous works are not concerned with the difference in the degree of difficulty, but only with the similarity of the fields of study. For example, if we only considered major similarity, Medicine and Nursing would be grouped together. However, Medicine requires a much higher grade for students to be called to fill a seat. Therefore, there may be students who choose Nursing over Medicine because they are more likely to enter this major.

[16]For example, Montmarquette et al. (2002) and Arcidiacono (2004) consider only 4 major groups.

Table 3 shows descriptive statistics of the different major groups. The first row shows the average minimum grade required to be called to fill a seat in a major of that group. A higher minimum grade indicates that it is in general more difficult to enter a major of that group. We can see that Health and Biological Sciences, and Engineering and Architecture are the two most difficult groups. Technologies and Exact Sciences, on the other hand, are the least difficult groups. The second row (candidates) shows the number and percentage of students who selected a major within that group as their first choice. The third row (called) shows the number and percentage of candidates who are called to fill a seat in their first choice major. Finally, the fourth row (enrolled) shows the number and percentage of students that decide to enroll in their first choice major when offered the chance.

[TABLE 3 ABOUT HERE.]

Table 3 shows there are differences in the proportion of men and women choosing the different majors, but also in the proportion of men and women who are called to enter a major. For example, women represent 22.37% of candidates for a seat in Engineering and Architecture, but represent only 18.94% of students called to fill a seat. This difference in shares means that men are offered seats in a higher proportion than women.

The dependent variables used in the estimations are the choice of major (first equation), and the outcome of the entry process (second equation). As explained in the previous section, the entry variable measures whether the candidate was effectively called to fill a seat in her first choice major. Table 4 shows the list of independent variables, and indicates the group of variables that will appear in each equation. Most variables will appear in both equations, but some of them will appear only in the equation determining the choice of major or in the equation determining the probability of entry. The table also shows which categories will be the reference categories in the estimations.

[TABLE 4 ABOUT HERE.]

The main explanatory variable is gender. The estimations also include interactions between gender and several variables. Other individual characteristics are represented by race, age and work status.

Education variables are also very important. There is currently an intense debate in Brazil on whether students coming out of public schools have a lower chance of entering public universities because of the low quality of primary and secondary public education. Also, students of technical schools are considered to be better in mathematics and related subjects, which may affect their probability of choosing and entering the different majors. Finally, the estimations also include a variable which indicates if the student is already enrolled in another major in University of Campinas or another university.

The most important socioeconomic variable is income. When students are surveyed, they are only asked in which income segment their family income lies. Therefore, we do not have actual income, but a categorical variable indicating the income of the family relative to the minimum wage. Allegedly, higher income should imply a higher probability of entering any major, but it is more difficult to conjecture what should be the effect on major choice. Also, poorer students are exempt from paying the registration fee, so this can also be used as an indicative of family wealth. Other important variables are the ones describing the occupation and education of father and mother.

It is always desirable to have exclusion variables to help with identification. Some variables only affect the probability of entry. For example, students with a high ENEM grade may add points to the vestibular score.[17] Also, if the student took a pre-vestibular course, she is likely to perform better in the vestibular exam. Finally, the entry equation also includes an interaction between ENEM and gender, to test whether higher ENEM grades have a differential impact on men and women. All these variables will influence the probability of entering the major the student chooses, but will not affect preferences.

Likewise, variables describing the reasons why the student chose the major and the University of Campinas will affect the equation determining major choice, but will not affect the probability of entering a given major. The choice equation also includes interactions between gender and work, secondary school variables,

[17]ENEM is a voluntary exam which students can take after finishing secondary school. Students who have taken the ENEM exam increase their final vestibular exam score, only if the ENEM grade is higher than the unadjusted vestibular score. If the student did not take the ENEM test, or if the ENEM grade is below the unadjusted vestibular score, then the vestibular score is not changed. Therefore, taking the ENEM exam may be beneficial (if the student gets a high score), but is never harmful for students.

other major, registration fee and the variables showing the reasons for choosing the major and the university.[18]

Table 5 shows summary statistics for the independent variables. Columns 1 and 2 show the average value of each variable for men and women, and column 3 shows the sample average. For categorical variables, the average is the proportion of individuals for whom the variable is equal to 1. Interestingly, we see important differences between men and women with respect to their individual characteristics, socioeconomic variables and education variables. This shows why it is important to control for all these characteristics in the regressions, when trying to elucidate the effect of gender on preferences and probabilities of entry.

[TABLE 5 ABOUT HERE.]

5. Results of the estimations

In this section, I discuss the estimation results for the models presented in Section 2. I start with the benchmark model (Model I), and then proceed to analyze the model with correlations (Model II).

5.1. Benchmark model. As explained in Section 2, the benchmark model is estimated in two stages. First, I estimate the parameters of the entry equation, and with the resulting parameters, I calculate the probabilities of entering the different majors for each individual. Second, I estimate a multinomial logit model for the choice of major, using entrance probabilities estimated in the first stage.

Table 6 shows the estimated coefficients for the entry equation (equation 46). Table 7 shows the corresponding average marginal effects (in percentage terms), which are calculated as the average of the marginal effects of all individuals. For Gender and ENEM, the average marginal effect includes the effect of the interaction, so the sign of the marginal effects may differ from the sign of the coefficients.

[18] I have also estimated a model with more interactions, and found interactions to be generally non significant in the entry equation (besides the interaction between gender and ENEM). In contrast, many interactions were significant in the choice equation. The unrestricted model had too many coefficients (each interaction increases the number of coefficients to be estimated by 9), which reduced the global significance of the model. For the purposes of this paper, therefore, I will only include in the estimations the interactions which were significant in the unrestricted model.

For all other variables, the sign of the coefficients will always coincide with the sign of the marginal effects.

[TABLE 6 ABOUT HERE.]

[TABLE 7 ABOUT HERE.]

With respect to the effect of gender on the probabilities of entry, a positive (negative) sign would indicate that males (females) have a greater probability of entering a given major group. The coefficient of gender is positive and significant for 6 groups, and it is non significant for 3 groups. Given that the estimations include an interaction between Gender and ENEM, this result only means that men have a higher probability for entering 6 major groups considering students with ENEM equal to 0. To perform a complete analysis of the effects of Gender on entrance probabilities, we have to study the sign and significance of the coefficients of ENEM and the interactions between ENEM and Gender.

The coefficient of ENEM is positive and significant for all groups. The coefficient for the interaction is negative and significant for 7 major groups, but is always smaller in absolute value than the ENEM coefficient. Therefore, a higher ENEM grade implies a higher average probability of entering all majors for both men and women, but for 7 groups of majors, a higher ENEM grade has a greater impact on women's probability of entry.

The first row of Table 7 shows the average marginal effects of Gender considering all students. We can see that the marginal effect is negative for 6 major groups, and is non significant for 3 groups. For example, holding other personal characteristics constant, women have on average a 7.80 percentage points higher probability of entering a major in Other Health and Biological Sciences, and a 4.78 percentage points higher probability of entering a Technologies major.

However, given that the interaction between ENEM and Gender is significant, the average marginal effect of gender on entrance probabilities depends on the ENEM grade. Table 8 shows the marginal effect of gender on entrance probabilities for groups of students with different ENEM. Given that for most majors ENEM coefficients for females are larger than for males, for a large enough ENEM this will affect the sign of the average marginal effect of gender conditional on ENEM.

In the case of Engineering and Architecture, for example, the average marginal effect is positive for low ENEM grades, is non significant for intermediate ENEM grades, and is negative for high ENEM grades. In the case of Health and Biological Sciences, on the other hand, the effect is non significant for most ENEM groups, and is negative and significant only for the group of students with highest ENEM grade.

[TABLE 8 ABOUT HERE.]

Going back to Table 7, White is significant only for 2 major groups, Arts and Health and Biological Sciences, and in both cases, the coefficient and marginal effect is positive. The fact that White is non significant for many groups may in part be due to the PAAIS affirmative action program, which gives additional points to black and aboriginal students, and may be thus counteracting any advantage White students may have. Work is significant and negative for 4 major groups (working implies a lower probability of entering university), and significant and positive for 1 major group.

The coefficients of the age variables are significant and negative for 7 major groups, and are positive for only one group, Health and Biological Sciences. For most major groups, the marginal effects decrease in absolute value as age increases. This result is surprising, because it means that for most majors, getting older has a positive effect on the probability of entry, which may be due to two effects. First, many students try to enter the university several years before succeeding. These students may have a higher chance of entering as time goes by because they become more experienced. Second, there may be a selection process, through which older students who are still trying to enter the university are the most constant and hard-working students. In the case of Health and Biological Sciences, on the other hand, the coefficients of the age categories are positive and marginal effects decrease with age, which means that younger students have an advantage on average to enter this major group.

The coefficient of Primary School Private is significant and positive for 5 major groups. The largest marginal effects are those corresponding to Humanities (4.90 percentage points) and Exact Sciences (4.72 percentage points). The coefficient of Secondary School Private is significant and negative for 4 major groups, and

significant and positive for 1 major group. The coefficient of Secondary School Mixed (students who attended both public and private schools) is significant and negative for 4 major groups. For Engineering and Architecture; Natural and Earth Sciences; and Social Sciences, coming from a private secondary school implies a decrease of 4 to 5 percentage points in the probability of entry. This surprising finding may in part be due to the PAAIS affirmative action program. Students who only attended public schools for their secondary education receive extra points in the vestibular exam, which may more than compensate the positive effects that attending private secondary schools could have on the probability of entering university.

The effect of attending a technical school is positive for 5 major groups, and the largest effect is on Humanities (7.8 percentage points). Surprisingly, the effect on Engineering and Architecture is negative. Nevertheless, it is important to remark that the sign and significance of the effects changes in the estimations corresponding to Model II. For example, in Model II, the effect on Engineering and Architecture will become non significant.

Students who are already enrolled in another major, at University of Campinas or another university, have in general a higher probability of accessing another major. For example, being enrolled in another major increases the probability of entering Humanities in 16.40 percentage points, and increases the probability of entering Exact Sciences in 15.26 percentage points. This result shows that having some experience in higher education has a positive impact on the possibilities of entering another major.

Being exempt of the registration fee indicates that the student comes from a poorer socioeconomic background. For almost all major groups, Registration Fee has a negative effect on entrance probabilities. Therefore students coming from a poorer economic background have in general a lower probability of entering university. The exception is Health and Biological Sciences, for which being exempt of the registration fee has no impact on entrance probabilities.

Finally, preparing for the vestibular exam in a private academy increases the probability of entering 7 majors. Interestingly, the preparation course has a greater effect on majors with a lower minimum grade, like Technologies, Exact Sciences,

and Other Health and Biological Sciences. For example, preparing for the vestibular exam increases the probability of entering a major in Other Health and Biological Sciences in 4.28 percentage points, and increases the probability of entering Technologies in 3.95 percentage points.

Next, I discuss the estimation results for the parameters of the choice equations (equation 44). Table 9 shows the estimated coefficients and Table 10 shows the corresponding average marginal effects. Marginal effects are calculated as the average of the individual marginal effects, and are shown in percentage terms. As with any polychotomous choice model, the sign of the coefficients may not coincide with the sign of the marginal effects because we have to consider the effect of a variable on the utility of one alternative, in comparison with the effect on the utility of the other alternatives. Moreover, the variables which are included in both equations have a double effect on the probability of choosing a major: on one hand, they affect the utility of entering the different majors, but at the same time they affect the probability of entering the different majors, which also affects expected utility. For these reasons, it is more useful to perform the analysis in terms of marginal effects.

[TABLE 9 ABOUT HERE.]

[TABLE 10 ABOUT HERE.]

After controlling for other individual characteristics, and taking into account the effect of the interactions, males have on average a greater probability of choosing mathematically-oriented majors, like Technologies, Exact Sciences, and Engineering and Architecture. Men also have a greater probability of choosing Social Sciences. Women, on the other hand, have a greater probability of choosing Health and Biological Sciences; Other Health and Biological Sciences; Natural and Earth Sciences; and Arts. These findings are consistent with those of the previous literature, and show that the Brazilian case exhibits similar patterns to those found in other countries.

With respect to the magnitude of the effects, the largest effects of gender are on Engineering and Architecture, and Health and Biological Sciences. In particular, men have a 13.86 percentage points higher probability of choosing Engineering

and Architecture, and women have a 10.38 percentage points higher probability of choosing Health and Biological Sciences.

The marginal effects of White and Work are generally significant. The largest effect of White is on the probability of choosing a major in Engineering and Architecture: white students have a 1.42 percentage points lower probability of choosing a major in this group. With respect to Work, students who work have a 11.39 percentage points lower probability of choosing Health and Biological Sciences, and a 16.13 percentage points higher probability of choosing Engineering and Architecture.

Students who went to a technical secondary school have a 21.12 percentage points higher probability of choosing Engineering and Architecture, and have a lower probability of choosing all other major groups. Being enrolled in another major is also generally significant. Students who are enrolled in another major, in the same or another university, have a 14.59 percentage points higher probability of choosing Engineering and Architecture, and a 5.41 percentage points lower probability of choosing Health and Biological Sciences.

Students who are exempt from paying the registration fee are more likely to choose Engineering and Architecture, Humanities, and Other Health and Biological Sciences, and are less likely to choose Technologies, Natural and Earth Sciences, Arts, Social Sciences and Health and Biological Sciences.

The coefficients of the variables indicating the reasons for choosing the major and the university are also significant. It is interesting to comment the results for some major groups. For example, choosing a major for job market reasons implies a decrease of 3.32 percentage points in the probability of choosing an Arts major and a decrease of 4.02 percentage points in the probability of choosing Health and Biological Sciences. Likewise, choosing a major for its social contribution implies an increase of 13.84 percentage points in the probability of choosing Health and Biological Sciences, and a decrease of 12.07 percentage points in the probability of choosing Engineering and Architecture.

5.2. Model with correlations. As explained in Section 2.2, Maximum Simulated Likelihood Estimators (MSLE) will be consistent if R grows at a rate larger or equal to $\sqrt{N}$, where N is the number of individuals. As a consequence, the number of computations increases at a rate of $N^{3/2}$, which makes it difficult to use a large sample. Therefore, for the model with correlations I will use the sample

corresponding to the year 2008.[19] This sample has 39,494 observations. I will use $R = 200$ draws for each individual, which is larger than $\sqrt{N}$.[20]

Table 11 shows the coefficients of the entry equations for Model II, and Table 12 shows the corresponding average simulated marginal effects. Reported marginal effects are the average of the marginal effects calculated for each individual and each draw, and are shown in percentage terms.

[TABLE 11 ABOUT HERE.]

[TABLE 12 ABOUT HERE.]

The coefficient of Gender is positive for 4 major groups, negative for 4 major groups, and non significant for 1 major group. The coefficient of the interaction between gender and ENEM is positive for 3 major groups, negative for 4 groups and non significant for 2 groups. As in Model I, whenever the coefficient of the interaction is negative, it is smaller in absolute value than the coefficient of ENEM, which means that a higher ENEM grade implies a higher average probability of entry for both men and women. Unlike Model I, however, the coefficient of the interaction is positive for some major groups, which means that for some majors, a higher ENEM grade has a larger impact on men in comparison with women.

As in the previous model, the presence of an interaction between Gender and ENEM implies that the average marginal effect of gender will vary depending on the ENEM grade of the group of students considered. Table 13 shows the marginal effect of gender for groups of students with different ENEM. Consider

[19] In order to determine the effects of using a smaller sample, in Appendix A, I include the estimation results for Model I using the sample corresponding to year 2008. Comparing the estimations of Model I with the full sample and the reduced sample, we can see that the differences in the effects of gender on choices are minimal. In the entry equation, there is a loss of significance of the gender effect for two major groups, but the sign and significance of the other groups remains unchanged.

[20] It may be argued that part of the differences in the results of Model I and II is caused by the difference in the estimation procedures. Specifically, Model II is estimated by MSL which involves simulating the expected value of the probabilities in equations 50 and 51. However, if the correlation between the estimations turns out to be non significant, then the expected value of these probabilities will be equal to the probabilities given by equations 47 and 48. Given the large number of draws used in the simulations, the simulated expected probabilities will be a good approximation of the true expected values. Therefore, the differences in the estimations will not be caused by the difference between ML and MSL.

first the two major groups with highest average minimum grades. For Engineering and Architecture, the average effect considering all students is positive, and is also positive for most ENEM groups. However, the average marginal effect becomes negative for the top ENEM category, which is due to the fact that the coefficient of the interaction between ENEM and Gender is negative for this group. Health and Biological Sciences shows a very different pattern. The average effect considering all students is negative, and is also negative for most ENEM groups (all groups but the top group, where it becomes non significant).

[TABLE 13 ABOUT HERE.]

If we now consider the major group with the lowest average minimum grade (Technologies) we can see that the marginal effect is negative for low ENEM grades and is non significant for higher ENEM grades. On the other hand, for Exact Sciences the average effect considering all students is non significant, and the effect is negative for students with ENEM equal to zero, and positive for the top ENEM groups. Similar analyses can be performed for other major groups, which lead to the conclusion that the relation between the average marginal effect of gender and the ENEM grade depends on the particular major group under analysis.

Going back to Table 12, White is significant for 4 major groups. The largest effect is on the probability of entering Exact Sciences (2.27 percentage points), Other Health and Biological Sciences (2.20 percentage points), and Arts (2.18 percentage points). Interestingly, Work is significant only for two major groups, Engineering and Architecture, and Health and Biological Sciences, and is negative in both cases. Therefore, students who work have a lower chance of entering the most demanding majors.

With respect to the coefficients of the age variables, there are 4 major groups which exhibit a similar pattern. For Natural and Earth Sciences, Social Sciences, Humanities, and Other Health and Biological Sciences, the marginal effects are negative and decreasing in absolute value as age increases. This means that older students have an advantage to enter a major in these groups. The possible reasons behind this result have already been discussed in the previous section.

Primary education variables lose significance in comparison with Model I. Primary School Private only has a positive effect for Humanities, and Primary School

Mixed has a positive effect on Engineering and Architecture; and Health and Biological Sciences.

Secondary School Private has a negative effect on the probability of entering the two most difficult majors (Engineering and Architecture; and Health and Biological Sciences). For these majors, the positive effect that private secondary schools may have is completely overcome by the extra points students get with the PAAIS program. Finally, the effect of Secondary School Mixed is negative for 2 major groups, Social Sciences and Health and Biological Sciences.

The effect of Secondary School Technical is positive and significant for Natural and Earth Sciences, and is negative for Arts and Health and Biological Sciences. This result is surprising because it means that attending a technical secondary school does not imply a higher probability of entry into technical and math-oriented majors.

The effect of being enrolled in another major is positive for all major groups except Arts, which means that having some experience in University increases the chances of entering a second career for most majors. Interestingly, the effect is larger for easier majors, and is smaller for the two most difficult majors (Engineering and Architecture, and Health and Biological Sciences).

The effect of Registration Fee is negative for all major groups except Arts. This result means that poorer students have on average a lower probability of entering most majors. As in Model I, Preparation Course is positive for 7 major groups. Finally, ENEM maintains sign and significance, which means that having a higher ENEM grade increases the probability of entering all majors.

Next, I analyze the estimation results for the parameters of the choice equations. Table 14 shows the estimated coefficients and Table 15 shows the corresponding simulated average marginal effects. In addition to the variables shown in the previous section, Table 14 shows the estimated coefficient and standard deviation of σ. The marginal effects are the average of the simulated effects corresponding to 200 draws for each individual, and show the average effects on the probability of choosing a major group in percentage terms.

[TABLE 14 ABOUT HERE.]

[TABLE 15 ABOUT HERE.]

Before analyzing the sign and significance of the effects, it is important to comment on the coefficient σ, which is significant. This means that the errors of the choice and entry equations are correlated: students who get a larger preference shock for some major tend to have a higher entry shock for that major as well. Econometric models that do not take this correlation into account will produce biased estimators, and therefore it is important to consider correlated errors in the econometric design.

In comparison with Model I, the sign of the average marginal effects of Gender on choice probabilities remains unchanged for all groups except for Social Sciences, for which the marginal effect is now non significant. The magnitudes of the effects are also similar to the previous case, except in the cases of Engineering and Health and Biological Sciences, for which they increase in absolute value. According to Model II, men have on average a 24.14 percentage points higher probability of choosing Engineering, controlling for other individual characteristics. Likewise, women have on average a 16.95 percentage points higher probability of choosing Health and Biological Sciences.

White has an effect on 4 major groups, and the largest effect is on Engineering and Architecture. Working decreases the probability of choosing Health and Biological Sciences in 12.05 percentage points, and increases the probability of choosing all other majors except Natural and Earth Sciences and Other Health and Biological Sciences, for which the effect is non significant.

Attending a technical secondary school implies a higher probability of choosing math-related majors (Technologies, Exact Sciences, and Engineering and Architecture), and Natural and Earth Sciences, and implies a lower probability of choosing Arts, Social Sciences, Health and Biological Sciences and Other Health and Biological Sciences. The largest effects are on Engineering and Architecture (7.78 percentage points), and Health and Biological Sciences (-7.77 percentage points). According to Model II, then, attending a technical secondary school does not affect the probability of entering mathematically oriented majors, but it does affect the probability of choosing these majors.

Being enrolled in another major also affects major choice. The two largest effects are on Health and Biological Sciences, and Social Sciences. Being enrolled in another major decreases the probability of choosing Health and Biological Sciences

in 7.81 percentage points, and increases the probability of choosing Social Sciences in 4.26 percentage points.

As in the previous model, students who are exempt from paying the registration fee are less likely to choose Engineering and Architecture, and Health and Biological Sciences, and more likely to choose Exact Sciences and Humanities. Therefore, the conclusions of the analysis of Model I still hold: poorer students tend to avoid choosing harder majors. In particular, being exempt from the registration fee implies a decrease of 15.54 percentage points in the probability of choosing Health and Biological Sciences, and a decrease of 10.98 percentage points in the probability of choosing Engineering and Architecture, which are the two most demanding major groups, according to Table 3.

Finally, the analysis of the reasons for choosing major and university lead to similar conclusions as before. Choosing a major for job market reasons implies a decrease of 3.54 percentage points in the probability of choosing Arts, a decrease of 7.39 percentage points in the probability of choosing Health and Biological Sciences, and an increase of 4.11 percentage points in the probability of choosing Engineering and Architecture. Likewise, choosing a major for its social contribution implies an increase of 21.26 percentage points in the probability of choosing Health and Biological Sciences, and a decrease of 18.24 percentage points in the probability of choosing an Engineering major.

5.3. Gender differences in entrance probabilities and preferences. According to the model presented in Section 2, students choose majors by comparing expected utilities ($p_{ij}\, u_{ij}$). As a consequence, gender affects major choice in two ways: (i) through its effect on entrance probabilities (p_{ij}), and (ii) through its effect on preferences (u_{ij}).

The marginal effects presented in the previous section were constructed taking into account both effects. In this section, I try to separate the two effects, to see what part of the difference in gender choices are generated by differences in the probability of entering the different majors, and what part is generated by differences in preferences.

Specifically, I perform two simulations. First, I simulate women's choices using male entrance probabilities (i.e. setting Gender equal to 1 in the entry equation, and equal to 0 in the choice equation), and compare them with men's choices (setting Gender equal to 1 in both equations). Then, I simulate men's choices using

female entrance probabilities, and compare them with women's choices. Table 16 presents the results of the simulations for Model II, as well as choice probabilities calculated with own-gender entrance probabilities.

[TABLE 16 ABOUT HERE.]

As expected, changing the entrance probabilities used to calculate expected utility has an effect on the probabilities of choosing the different majors. For example, using female entrance probabilities, women's average probability of choosing Engineering and Architecture is 12.72%, but when we simulate women's choices using male entrance probabilities, this probability increases to 21.11%. Likewise, using male entrance probabilities, men's average probability of choosing Health and Biological Sciences is 22.41%, but when we simulate men's choices using female entrance probabilities, this probability increases to 30.07%. Therefore, it is clear that gender differences in entrance probabilities affect major choice. In particular, men have on average a higher probability of entering Engineering and Architecture, and a lower probability of entering Health and Biological Sciences (see Table 12). Therefore, men will choose Engineering and Architecture majors in a higher proportion than they would choose them if this difference in entrance probabilities did not exist. Likewise, women will choose Health and Biological Sciences majors in a higher proportion than they would choose them if this difference in entrance probabilities did not exist.

Table 16 also shows that the sign of the gender differences in choice probabilities is the same for both simulations, except in the case of Social Sciences. Moreover, the magnitudes of the differences are similar. For example, in the case of Engineering and Architecture, simulated gender differences in choice probabilities are 16.24 or 17.79 percentage points, depending on the simulation.

For most majors, simulated gender differences in choice probabilities are very similar to gender differences calculated using own-gender probabilities. Therefore, for these majors, differences in preferences explain most of the gender difference in major choice. Nevertheless, there are two important exceptions, which are precisely the two most difficult majors. Using own-gender entrance probabilities, men have on average a 26.28 percentage points higher probability of choosing Engineering and Architecture, but when we simulate choices controlling for gender

differences in entrance probabilities, men have a 16.24 or 17.79 percentage points higher probability. Therefore, for Engineering and Architecture, there is a substantial part of the difference in choices which is explained by gender differences in the probability of entry. The same can be said about Health and Biological Sciences.

5.4. Interaction between gender and other explanatory variables. Given the presence of interactions between gender and other variables, the marginal effect of gender may differ for groups of students with different characteristics. For example, gender differences in choice probabilities may be smaller or greater for students who attended public secondary schools, in comparison with students who attended private secondary schools. Table 17 shows simulated average marginal effects for different groups of students for Model II.

[TABLE 17 ABOUT HERE.]

Table 17 shows there are significant differences in the effect of gender depending on the group of students under analysis. For example, comparing the effect of gender for working and non-working students, we can see that in the cases of Engineering and Architecture, Natural and Earth Sciences, Humanities, and Health and Biological Sciences, the difference in choice probabilities between men and women is larger (in absolute value) for students who work than for students who do not work. Nevertheless, this pattern is not uniform across majors. In the cases of Technologies, Exact Sciences, Arts, Social Sciences, and Other Health and Biological Sciences, the gender difference in choice probabilities is smaller for students who work, in comparison with students who do not work.

Similar analyses can be performed for other variables. In particular, it is interesting to examine the effects of secondary education on gender differences. For example, men have a 25.56 percentage points higher probability of choosing Engineering and Architecture if we consider the group of students who attended private secondary schools, but the difference reduces to 21.34 percentage points for students who attended public secondary schools. Likewise, men have a 19.82 percentage points lower probability of choosing Health and Biological Sciences in the group of students who attended private secondary schools, but the difference reduces to 11.13 percentage points for students who attended public secondary

schools. Therefore, private secondary education leads to higher differences in the choices of men and women in the two most demanding major groups. Nevertheless, there are also major groups for which gender differences are larger in the group of students who attended public secondary schools (e.g. Other Health and Biological Sciences).

6. Conclusion

Gender differences in major choice have triggered an extensive literature trying to decipher the reasons for the existence of such differences. The Brazilian case is interesting, because in most public universities students choose a major before taking a major-specific exam which determines whether they can enter the major of their choice. This contrasts with the college entrance process in most other countries (including the US), where students are first allowed entry into university and then have to choose their preferred major. The singular characteristics of the Brazilian case allow us to test whether differences in choices are due to differences in the probabilities of entry or differences in the utility associated with the different majors.

I have presented two econometric models, and estimated them using data from the University of Campinas, a prestigious public university dependent of the State of São Paulo. The first model imposes independence between preference and entry shocks, but can be estimated with standard econometric software. The second model relaxes the independence assumption, but becomes harder to estimate, and I have to resort to a Maximum Simulated Likelihood approach.

After estimating the second model, I find that the correlation between the errors of the two equations is positive and significant: students who get a larger preference shock for some major tend to have a higher entry shock for that major as well. The significance of this coefficient means that the model without correlations will produce biased estimators. Therefore, it is important to consider correlated errors in the econometric design.

With respect to the effect of gender on entrance probabilities, there are several interesting findings. First, the average gender effect on entrance probabilities is positive for some majors and negative for other majors. Second, the effect of gender on entrance probabilities depends on the ENEM grade. Nevertheless, it is difficult

to generalize on the nature of the relation between gender effects and ENEM, as it will generally depend on the specific major group under consideration.

In addition to gender, entrance probabilities are affected by other variables. ENEM has a positive effect on entrance probabilities for both men and women. Students who are already enrolled in another major have a higher probability of entering all major groups except Arts, for which the effect is non significant. Students who are exempted of the registration fee (which indicates that the student comes from a poorer family) have a lower probability of entering all majors except Arts, for which the effect is non significant.

An important issue being discussed in Brazil is what is the effect of private vs. public education on the possibilities of accessing higher education. It is generally argued that students of private schools receive a better education, which gives them an advantage for entering college. I find that the effects of attending a primary private school are generally non significant. I also find that students who attended private secondary schools have a lower probability of entering the most demanding majors (Engineering and Architecture, and Health and Biological Sciences), and have a higher probability of entering Exact Sciences and Other Health and Biological Sciences. In the case of Engineering and Architecture, and Health and Biological Sciences, the negative sign of the coefficient may be partly due to the PAAIS affirmative action program, which gives additional points to students who attended only public secondary schools.

With respect to the effects of gender on major choice, I find that men have on average a higher probability of choosing mathematically oriented majors (Technologies, Exact Sciences and Engineering and Architecture), and women have on average a higher probability of choosing Natural and Earth Sciences, Arts, Humanities, Health and Biological Sciences and Other Health and Biological Sciences. The average effect of gender on the probability of choosing Social Sciences is non significant.

In order to determine if gender differences in major choice are caused by differences in preferences or probability of entry, I simulate women choices with male probabilities of entry, and men choices with female probabilities of entry. I find that preferences account for most of the difference in choices in majors with low or medium minimum required grades. In the most demanding majors (Engineering and Architecture, and Health and Biological Sciences), on the other hand, a large

part of the difference in major choice is explained by differences in the probability of entry.

Finally, I find that the effect of gender on major choice depends on education, socioeconomic variables and family background. For example, for Engineering and Architecture, and Health and Biological Sciences, the difference between men and women is larger among students who attended private schools, in comparison with students who attended public schools. Therefore, for these two major groups, private secondary education leads to larger differences between men and women. Nevertheless, there are also major groups for which gender differences are larger in the group of students who attended public secondary schools (e.g. Other Health and Biological Sciences).

TABLE A1. Marginal effects on the probability of entry, year 2008 (Model I)

	Technologies	Exact Sciences	Engineering and Arch.	Natural and Earth Sc.	Arts	Social Sciences	Humanities	Health and Biol. Sc.	Other Health and Biol. Sc.
Gender	-11.78 ***	-1.96	-0.61	-0.49	-1.03	-1.68 *	-1.23	-0.82 ***	-7.60 ***
	(3.69)	(2.52)	(0.66)	(1.65)	(2.46)	(0.86)	(3.35)	(0.31)	(2.30)
White	1.63	4.36 *	0.20	0.23	4.25	-0.39	0.21	0.21	0.99
	(3.79)	(2.29)	(0.60)	(1.64)	(2.78)	(1.06)	(3.11)	(0.30)	(2.06)
Work	-2.36	-2.13	-1.42 *	-1.19	5.34 *	-1.38	-1.75	0.84	-4.42 *
	(4.18)	(2.63)	(0.85)	(2.00)	(2.94)	(1.28)	(2.99)	(0.68)	(2.48)
Age1	7.88	-19.46 ***	-0.65	-6.27	-21.01 ***	-7.55 **	-17.22 ***	1.26 *	-20.43 ***
	(8.91)	(6.02)	(2.70)	(4.66)	(8.11)	(3.80)	(6.08)	(0.64)	(5.61)
Age2	0.82	-18.26 ***	-0.60	-3.16	-19.33 **	-5.88 *	-8.86 *	1.42 ***	-13.58 ***
	(6.92)	(5.36)	(2.59)	(4.13)	(7.66)	(3.55)	(5.02)	(0.48)	(5.06)
Age3	10.72	-11.78 **	-0.80	-4.57	-14.65 **	-5.61	-2.94	1.24 ***	-8.15
	(6.58)	(5.19)	(2.55)	(4.06)	(7.47)	(3.49)	(4.87)	(0.46)	(4.97)
Prim Sch Priv	-1.82	1.76	2.70 ***	-0.90	0.38	1.85	14.52 ***	0.49	6.25 **
	(5.41)	(3.34)	(0.89)	(2.14)	(3.82)	(1.40)	(4.18)	(0.45)	(2.45)
Prim Sch Mixed	0.37	0.25	1.33	1.37	0.19	0.41	7.02	0.05	2.98
	(5.36)	(3.34)	(0.93)	(2.32)	(3.91)	(1.48)	(4.32)	(0.49)	(2.58)
Sec Sch Priv	3.75	6.65 **	-5.40 ***	-3.94 *	0.18	-4.23 ***	1.61	-1.90 ***	3.76
	(5.31)	(2.85)	(0.99)	(2.10)	(3.59)	(1.54)	(4.37)	(0.60)	(2.40)
Sec Sch Mixed	-16.21 **	1.95	-4.69 ***	-4.54	-2.43	-7.41 ***	-7.32	-3.00 ***	-1.67
	(7.95)	(4.21)	(1.63)	(3.31)	(4.69)	(2.14)	(5.57)	(0.72)	(3.73)
Sec Sch Tech	1.31	-3.14	-0.17	5.74 **	-2.94	2.54	9.07 *	1.80 **	-1.55
	(5.00)	(2.87)	(0.85)	(2.34)	(4.63)	(1.84)	(4.77)	(0.83)	(3.30)

TABLE A1. Marginal effects on the probability of entry, year 2008 (Model I, cont.)

	Technologies	Exact Sciences	Engineering and Arch.	Natural and Earth Sc.	Arts	Social Sciences	Humanities	Health and Biol. Sc.	Other Health and Biol. Sc.
Other Major	15.68 ***	17.78 ***	3.93 ***	12.54 ***	8.81 *	9.12 ***	18.17 ***	0.48	14.25 ***
	(5.20)	(3.90)	(1.35)	(3.53)	(4.83)	(2.05)	(4.50)	(0.58)	(4.07)
Reg Fee	-27.76 ***	-13.28 ***	-3.32 **	-6.23 **	-1.86	-9.35 ***	-9.89 ***	-0.89	-4.02
	(4.94)	(2.91)	(1.43)	(2.50)	(5.26)	(0.90)	(3.34)	(0.82)	(3.01)
Prep Course	6.11 *	1.74	1.92 ***	4.83 ***	4.64 *	1.72 *	3.68	0.40	3.14 *
	(3.50)	(2.14)	(0.52)	(1.45)	(2.39)	(0.92)	(2.72)	(0.32)	(1.89)
ENEM	0.27 ***	0.37 ***	0.62 ***	0.78 ***	0.31 ***	0.31 ***	0.40 ***	0.40 ***	0.63 ***
	(0.05)	(0.07)	(0.09)	(0.12)	(0.06)	(0.06)	(0.06)	(0.04)	(0.09)
Number of obs.	39,494								

* significant at 10%, ** significant at 5%, *** significant at 1%.

TABLE A2. Marginal effects on choice probabilities (Model I)

	Technologies	Exact Sciences	Engineering and Arch.	Natural and Earth Sc.	Arts	Social Sciences	Humanities	Health and Biol. Sc.	Other Health and Biol. Sc.
Gender	0.70 ***	3.87 ***	14.67 ***	-2.61 ***	-0.98 ***	1.14 ***	-2.86 ***	-10.12 ***	-3.82 ***
	(0.14)	(0.21)	(0.36)	(0.25)	(0.17)	(0.30)	(0.17)	(0.38)	(0.21)
White	-0.10	-0.25	-1.51 ***	0.55 **	-0.10	0.18	0.01	0.48	0.73 ***
	(0.15)	(0.24)	(0.43)	(0.27)	(0.21)	(0.35)	(0.18)	(0.42)	(0.23)
Work	-0.17	-2.17 ***	23.22 ***	-3.34 ***	-0.26	-3.82 ***	0.96 ***	-13.46 ***	-0.95 ***
	(0.16)	(0.23)	(0.72)	(0.29)	(0.36)	(0.44)	(0.27)	(0.64)	(0.30)
Prim Sch Priv	-0.44 **	0.00	0.45	0.13	-0.07	0.40	-0.77 **	2.36 ***	-2.07 ***
	(0.21)	(0.31)	(0.60)	(0.34)	(0.28)	(0.47)	(0.34)	(0.58)	(0.35)
Prim Sch Mixed	-0.46 **	0.38	-0.94	0.33	0.27	1.09 **	-0.74 **	1.37 **	-1.29 ***
	(0.21)	(0.33)	(0.65)	(0.37)	(0.30)	(0.51)	(0.31)	(0.63)	(0.37)
Sec Sch Priv	-6.15 ***	-5.30 ***	20.73 ***	-2.67 ***	-1.75 ***	-4.62 ***	-1.05	-0.05	0.85
	(1.30)	(1.00)	(1.61)	(0.53)	(0.56)	(0.87)	(0.80)	(0.18)	(0.54)
Sec Sch Mixed	-9.66 ***	-4.49 ***	16.83 ***	-1.79 ***	-1.04 *	-3.46 ***	-2.31 *	5.27 ***	0.66
	(2.46)	(1.60)	(2.20)	(0.66)	(0.62)	(1.10)	(1.21)	(0.80)	(0.67)
Sec Sch Tech	-0.56 ***	-2.29 ***	25.45 ***	-2.21 ***	-1.43 ***	-6.37 ***	-1.02 ***	-9.52 ***	-2.07 ***
	(0.19)	(0.28)	(0.74)	(0.35)	(0.26)	(0.43)	(0.26)	(0.72)	(0.34)
Other Major	-1.26 ***	-2.25 ***	21.98 ***	-3.70 ***	-2.05 ***	-1.52 **	-0.53 *	-8.27 ***	-2.40 ***
	(0.15)	(0.33)	(0.94)	(0.38)	(0.20)	(0.68)	(0.29)	(0.78)	(0.32)
Reg Fee	-0.74 ***	-1.88 ***	23.10 ***	-4.76 ***	-2.15 ***	-5.42 ***	8.81 ***	-17.88 ***	0.90
	(0.14)	(0.33)	(0.81)	(0.34)	(0.21)	(0.56)	(1.11)	(0.73)	(0.66)
Prof Father Non-manual	0.35 *	0.78 ***	0.24	0.69 **	0.47 **	-0.70	-0.07	-1.36 ***	-0.39
	(0.19)	(0.29)	(0.51)	(0.31)	(0.23)	(0.44)	(0.23)	(0.51)	(0.27)
Prof Father Manual	0.56 **	0.76 *	1.41	1.00 **	0.51	0.10	-0.44	-3.51 ***	-0.40
	(0.26)	(0.44)	(0.96)	(0.51)	(0.38)	(0.65)	(0.32)	(0.85)	(0.40)
Prof Father Other	0.42 *	1.05 ***	-0.37	1.94 ***	0.07	-0.59	0.18	-2.35 ***	-0.35
	(0.23)	(0.40)	(0.71)	(0.45)	(0.34)	(0.57)	(0.32)	(0.69)	(0.36)

TABLE A2. Marginal effects on choice probabilities (Model I, cont.)

	Technologies	Exact Sciences	Engineering and Arch.	Natural and Earth Sc.	Arts	Social Sciences	Humanities	Health and Biol. Sc.	Other Health and Biol. Sc.
Prof Mother Non-manual	-0.26	0.22	-0.01	0.36	0.36	-0.19	0.14	-1.33 ***	0.71 ***
	(0.19)	(0.27)	(0.47)	(0.32)	(0.22)	(0.39)	(0.24)	(0.48)	(0.27)
Prof Mother Manual	-0.13	-0.12	2.14 *	0.77	-0.69 *	-1.02	0.26	-1.17	-0.05
	(0.37)	(0.57)	(1.20)	(0.65)	(0.41)	(1.08)	(0.48)	(1.18)	(0.57)
Prof Mother Housewife	0.07	0.68 **	1.76 ***	0.75 **	-0.56 ***	-0.84 **	-0.19	-2.04 ***	0.37
	(0.22)	(0.30)	(0.53)	(0.35)	(0.21)	(0.43)	(0.26)	(0.52)	(0.29)
Prof Mother Other	-0.16	0.40	2.04 ***	0.97 **	-0.20	-0.62	0.02	-3.04 ***	0.58
	(0.23)	(0.40)	(0.71)	(0.47)	(0.30)	(0.58)	(0.32)	(0.69)	(0.39)
Rsn Maj Job Mkt	2.65 ***	1.69 ***	2.99 ***	4.13 ***	-3.32 ***	-0.11	-1.87 ***	-4.86 ***	-1.30 ***
	(0.34)	(0.45)	(0.75)	(0.55)	(0.30)	(0.58)	(0.26)	(0.68)	(0.37)
Rsn Maj Soc Cont	0.24	-2.81 ***	-12.52 ***	-0.34	-1.79 ***	2.07 ***	0.70 **	14.04 ***	0.42
	(0.21)	(0.28)	(0.61)	(0.40)	(0.26)	(0.53)	(0.27)	(0.70)	(0.34)
Rsn Maj Pers Real	-0.32 **	-0.35	-3.22 ***	0.09	0.22	-0.29	-0.41 **	3.27 ***	1.01 ***
	(0.13)	(0.25)	(0.43)	(0.28)	(0.20)	(0.35)	(0.19)	(0.43)	(0.25)
Rsn Maj Other	3.06 ***	0.47	-4.34 ***	2.16 ***	-1.32 ***	-0.18	0.44	-2.11 ***	1.83 ***
	(0.37)	(0.42)	(0.71)	(0.58)	(0.31)	(0.60)	(0.34)	(0.71)	(0.47)
Rsn Univ Free	0.32 *	-0.99 ***	-2.02 ***	-2.11 ***	-1.31 ***	1.42 ***	0.29	3.35 ***	1.06 ***
	(0.18)	(0.26)	(0.49)	(0.32)	(0.21)	(0.41)	(0.21)	(0.50)	(0.30)
Rsn Univ Rep	0.03	-0.89 ***	0.83 **	-1.37 ***	-1.72 ***	1.18 ***	0.63 ***	1.55 ***	-0.23
	(0.16)	(0.25)	(0.42)	(0.29)	(0.19)	(0.35)	(0.21)	(0.42)	(0.25)
Rsn Univ Other	0.71 ***	0.65 **	-0.05	-0.28	-1.47 ***	0.62	0.38 *	-0.32	-0.24
	(0.19)	(0.29)	(0.50)	(0.33)	(0.22)	(0.41)	(0.22)	(0.50)	(0.28)
Number of observations	39,494								

* significant at 10%, ** significant at 5%, *** significant at 1%.

Bibliography

Altonji, J. (1993): "The Demand for and Return to Education When Education Outcomes are Uncertain," *Journal of Labor Economics*, 11(1), 48.

Anderson, S., A. De Palma, and J. Thisse (1992): *Discrete Choice Theory of Product Differentiation*. MIT Press.

Arcidiacono, P. (2004): "Ability sorting and the returns to college major," *Journal of Econometrics*, 121(1-2), 343–375.

Ben-Akiva, M., and S. Lerman (1985): *Discrete choice analysis: theory and application to travel demand*. MIT press.

Cavalcanti, T., J. Guimarães, and B. Sampaio (2009): "Barriers to Skill Acquisition in Brazil: Evidences from a University Entrance Exam," Unpublished Manuscript.

Freeman, R. (1971): *The market for college-trained manpower: A study in the economics of career choice*. Harvard University Press.

Guimarães, J., and B. Sampaio (2007): "The Influence Of Family Background And Individual Characteristics On Entrance Tests Scores Of Brazilian University Students," Anais do XXXV Encontro Nacional de Economia 092, ANPEC - Associação Nacional dos Centros de Pósgraduação em Economia.

——— (2008): "Mind the Gap: Evidences from Gender Differences in Scores in Brazil," Anais do XXXVI Encontro Nacional de Economia 200807211527140, ANPEC - Associação Nacional dos Centros de Pósgraduação em Economia.

Mallar, C. (1977): "The estimation of simultaneous probability models," *Econometrica*, 45(7), 1717–1722.

McFadden, D. (1974): "Conditional logit analysis of qualitative choice behavior," in *Frontiers in econometrics*, ed. by P. Zarembka, pp. 105–142. Academic Press, New York.

McFadden, D., and K. Train (2000): "Mixed MNL models for discrete response," *Journal of Applied Econometrics*, 5, 447–470.

Montmarquette, C., K. Cannings, and S. Mahseredjian (2002): "How do young people choose college majors?," *Economics of Education Review*, 21(6), 543–556.

Pinheiro, L., N. de Oliveira Fontoura, A. C. Querino, A. Bonetti, and W. Rosa (2008): *Brasil–Retrato das Desigualdades: Gênero e Raça*. IPEA, SPM, UNIFEM, Brasilia, 3rd edn.

Train, K. (2003): *Discrete choice methods with simulation*. Cambridge University Press.

Turner, S., and W. Bowen (1999): "Choice of major: The changing (unchanging) gender gap," *Industrial and Labor Relations Review*, 52(2), 289–313.

Zafar, B. (2009): "College Major Choice and the Gender Gap," Federal Reserve Bank of New York, Staff Report No. 364.

TABLE 1. Candidates for entry between 2006 and 2008

	Candidates	Called			Enrolled		
		1st Choice	Other Choice	Total	1st Choice	Other Choice	Total
2006	40,162	4,180	720	4,900	2,389	317	2,706
%		85.31	14.69	100	88.29	11.71	100
2007	40,402	4,632	883	5,515	2,508	243	2,751
%		83.99	16.01	100	91.17	8.83	100
2008	39,494	4,354	549	4,903	2,518	178	2,696
%		88.80	11.20	100	93.40	6.60	100
Total	120,058	13,166	2,152	15,318	7,415	738	8,153
%		85.95	14.05	100	90.95	9.05	100

TABLE 2. Gender composition of candidates for entry between 2006 and 2008

	Candidates			Called			Enrolled		
	M	F	T	M	F	T	M	F	T
2006	20,232	19,930	40,162	2,417	1,763	4,180	1,362	1,027	2,389
%				11.95	8.85	10.41	6.73	5.15	5.95
2007	20,387	20,015	40,402	2,650	1,982	4,632	1,419	1,089	2,508
%				13.00	9.90	11.46	6.96	5.44	6.21
2008	19,652	19,842	39,494	2,373	1,981	4,354	1,381	1,137	2,518
%				12.08	9.98	11.02	7.03	5.73	6.38
Total	60,271	59,787	120,058	7,440	5,726	13,166	4,162	3,253	7,415
%				12.34	9.58	10.97	6.91	5.44	6.18

Rows marked with the percentage sign indicate the percentage of students called or enrolled in their first choice major over the total number of candidates.

TABLE 3. Descriptive statistics of major groups.

	Technologies	Exact Sciences	Engineering and Arch.	Natural and Earth Sc.	Arts	Social Sciences	Humanities	Health and Biol. Sc.	Other Health and Biol. Sc.	Total
Average Min. Grade	360.87	441.90	552.80	482.51	448.19	521.65	447.76	603.08	443.00	
Candidates	2,597	7,113	28,827	9,632	3,914	14,624	5,622	37,632	10,097	120,058
Male	1,639	5,148	22,378	4,118	1,679	7,494	1,455	13,593	2,767	60,271
%	63.11	72.37	77.63	42.75	42.9	51.24	25.88	36.12	27.40	50.20
Female	958	1,965	6,449	5,514	2,235	7,130	4,167	24,039	7,330	59,787
%	36.89	27.63	22.37	57.25	57.1	48.76	74.12	63.88	72.60	49.80
Called 1st ch	908	1,388	3,206	1,422	460	1,649	925	1,634	1,574	13,166
Male	578	1,059	2,573	711	201	903	307	682	426	7,440
%	63.66	76.3	80.26	50.00	43.70	54.76	33.19	41.74	27.06	56.51
Female	330	329	633	711	259	746	618	952	1,148	5,726
%	36.34	23.7	19.74	50.00	56.30	45.24	66.81	58.26	72.94	43.49
Enrolled 1st ch	638	820	1,642	935	380	815	603	702	880	7,415
Male	417	641	1,331	431	169	441	196	267	269	4,162
%	65.36	78.17	81.06	46.1	44.47	54.11	32.5	38.03	30.57	56.13
Female	221	179	311	504	211	374	407	435	611	3,253
%	34.64	21.83	18.94	53.9	55.53	45.89	67.5	61.97	69.43	43.87

TABLE 4. Variable definitions

Variable	Definition
Personal characteristics	
Gender	1 if male, 0 if female
White	1 if white, 0 otherwise
Work	1 if currently working, 0 if not working
Age1	1 if age is 17 or less
Age2	1 if age is between 18 and 19
Age3	1 if age is between 20 and 23
Age4 (*)	1 if age is 24 or more
Education variables	
Prim Sch Private	1 if attended only private primary schools
Prim Sch Public (*)	1 if attended only public primary schools
Prim Sch Mixed	1 if attended both private and public pr. sch.
Sec Sch Private	1 if attended only private secondary schools
Sec Sch Public (*)	1 if attended only public secondary schools
Sec Sch Mixed	1 if attended both private and public sec. sch.
Sec Sch Technical	1 if attended technical secondary school
Other Major	1 if already coursing another major
Socioeconomic factors	
Reg Fee	1 if exempt from paying registration fee
Income Low (*)	1 if family income is up to 5 minimum wages
Income Medium	1 if family inc. is between 5 and 15 min. wages
Income High	1 if family income is above 15 minimum wages
Educ Father None	1 if father has some primary school or none
Educ Father Prim	1 if father finished primary school
Educ Father Low Sec	1 if father finished low secondary school
Educ Father High Sec	1 if father finished high secondary school
Educ Father Uni (*)	1 if father finished university
Educ Mother None	1 if mother has some primary school or none
Educ Mother Prim	1 if mother finished primary school
Educ Mother Low Sec	1 if mother finished low secondary school
Educ Mother High Sec	1 if mother finished high secondary school
Educ Mother Uni (*)	1 if mother finished university

TABLE 4. Variable definitions (cont.)

Variable	Definition
Socioeconomic factors (cont.)	
Prof Father Professional (*)	1 if father is professional
Prof Father Non-manual	1 if father has job with non-manual tasks
Prof Father Manual	1 if father has job with manual tasks
Prof Father Other	1 if father has another kind of job
Prof Mother Professional (*)	1 if mother is professional
Prof Mother Non-manual	1 if mother has job with non-manual tasks
Prof Mother Manual	1 if mother has job with manual tasks
Prof Mother Housewife	1 if mother is a housewife
Prof Mother Other	1 if mother has another kind of job
Others	
Year	Vestibular exam year
Only in major choice equation	
Rsn Major Ability (*)	1 if chose major because of personal ability
Rsn Major Job Market	1 if chose major because of job market prospects
Rsn Major Soc Contrib	1 if chose major to contribute to society
Rsn Major Pers Realization	1 if chose major for personal realization
Rsn Major Other	1 if chose major for other reasons
Rsn Univ Best for course (*)	1 if chose Unicamp because is best for course
Rsn Univ Free	1 if chose Unicamp because it is free
Rsn Univ Reputation	1 if chose Unicamp for its reputation
Rsn Univ Other	1 if chose Unicamp for other reasons
Only in probability of entry equation	
Prep Course	1 if took a preparation course for vestibular exam
ENEM	ENEM test grade
ENEM * Gender	ENEM interacted with Gender

(*) Reference category in the estimations.

TABLE 5. Summary statistics

Variable	Male	Female	Total
Personal characteristics			
White	0.751	0.769	0.760
Work	0.203	0.147	0.175
Age1	0.165	0.172	0.168
Age2	0.534	0.554	0.544
Age3	0.222	0.218	0.220
Age4	0.079	0.056	0.067
Education variables			
Prim Sch Private	0.540	0.511	0.526
Prim Sch Public	0.297	0.338	0.317
Prim Sch Mixed	0.164	0.151	0.157
Sec Sch Private	0.650	0.640	0.645
Sec Sch Public	0.294	0.309	0.301
Sec Sch Mixed	0.056	0.051	0.053
Sec Sch Technical	0.112	0.065	0.089
Other Major	0.100	0.069	0.084
Socioeconomic factors			
Reg Fee	0.068	0.127	0.098
Income Low	0.264	0.323	0.293
Income Medium	0.470	0.446	0.458
Income High	0.266	0.231	0.249
Educ Father None	0.059	0.067	0.063
Educ Father Prim	0.072	0.087	0.079
Educ Father Low Sec	0.075	0.082	0.079
Educ Father High Sec	0.285	0.300	0.293
Educ Father Uni	0.508	0.463	0.486
Educ Mother None	0.048	0.054	0.051
Educ Mother Prim	0.074	0.086	0.080
Educ Mother Low Sec	0.081	0.090	0.086
Educ Mother High Sec	0.312	0.320	0.316
Educ Mother Uni	0.484	0.450	0.467

TABLE 5. Summary statistics (cont.)

Variable	Male	Female	Total
Socioeconomic factors (cont.)			
Prof Father Professional	0.487	0.453	0.470
Prof Father Non-manual	0.274	0.281	0.278
Prof Father Manual	0.105	0.118	0.111
Prof Father Other	0.134	0.148	0.141
Prof Mother Professional	0.320	0.296	0.308
Prof Mother Non-manual	0.261	0.269	0.265
Prof Mother Manual	0.043	0.043	0.043
Prof Mother Housewife	0.272	0.290	0.281
Prof Mother Other	0.105	0.102	0.104
Only in major choice equation			
Rsn Major Ability	0.541	0.485	0.513
Rsn Major Job Market	0.090	0.060	0.075
Rsn Major Soc Contrib	0.087	0.126	0.106
Rsn Major Pers Realization	0.204	0.259	0.232
Rsn Major Other	0.078	0.069	0.074
Rsn Univ Best for course	0.418	0.373	0.396
Rsn Univ Free	0.174	0.203	0.189
Rsn Univ Reputation	0.232	0.254	0.243
Rsn Univ Other	0.176	0.169	0.172
Only in probability of entry equation			
Prep Course	0.585	0.613	0.599
ENEM	89.987	83.208	86.540

For dummy variables, the mean is equal to the proportion of individuals with that characteristic.

TABLE 6. Coefficients of entry equations (Model I)

	Technologies	Exact Sciences	Engineering and Arch.	Natural and Earth Sc.	Arts	Social Sciences	Humanities	Health and Biol. Sc.	Other Health and Biol. Sc.
Gender	0.561 **	1.694 ***	2.382 **	4.482 ***	0.768	0.711	1.941 ***	1.105	2.081 ***
	(0.273)	(0.495)	(0.952)	(0.773)	(0.469)	(0.761)	(0.341)	(1.348)	(0.583)
White	0.123	0.085	0.046	-0.055	0.343 **	-0.013	0.114	0.189 ***	0.084
	(0.100)	(0.076)	(0.047)	(0.073)	(0.152)	(0.068)	(0.095)	(0.070)	(0.077)
Work	-0.016	-0.300 ***	-0.351 ***	-0.221 **	0.373 ***	-0.100	-0.071	0.023	-0.356 ***
	(0.108)	(0.087)	(0.074)	(0.096)	(0.121)	(0.087)	(0.095)	(0.130)	(0.096)
Age1	-0.603 ***	-0.866 ***	-0.298 *	-0.010	-1.158 ***	-0.530 ***	-0.781 ***	0.776 ***	-0.743 ***
	(0.224)	(0.164)	(0.174)	(0.197)	(0.278)	(0.195)	(0.183)	(0.202)	(0.197)
Age2	-0.346 **	-0.799 ***	-0.273 *	-0.041	-1.035 ***	-0.420 **	-0.645 ***	0.751 ***	-0.540 ***
	(0.173)	(0.140)	(0.165)	(0.173)	(0.244)	(0.173)	(0.143)	(0.176)	(0.168)
Age3	-0.083	-0.599 ***	-0.358 **	-0.067	-0.690 ***	-0.397 **	-0.396 ***	0.486 ***	-0.304 *
	(0.164)	(0.139)	(0.165)	(0.171)	(0.241)	(0.173)	(0.141)	(0.170)	(0.164)
Prim Sch Priv	-0.053	0.296 ***	0.479 ***	0.104	0.051	0.271 ***	0.283 **	0.129	0.213 **
	(0.145)	(0.101)	(0.074)	(0.099)	(0.177)	(0.095)	(0.130)	(0.102)	(0.093)
Prim Sch Mixed	0.057	0.085	0.261 ***	0.055	0.131	0.079	0.105	-0.018	0.104
	(0.146)	(0.104)	(0.080)	(0.104)	(0.178)	(0.103)	(0.132)	(0.111)	(0.100)
Sec Sch Priv	0.162	0.266 ***	-0.535 ***	-0.402 ***	-0.008	-0.499 ***	0.152	-0.646 ***	0.096
	(0.136)	(0.092)	(0.062)	(0.092)	(0.163)	(0.086)	(0.125)	(0.096)	(0.093)
Sec Sch Mixed	-0.160	0.118	-0.535 ***	-0.378 **	-0.308	-0.773 ***	0.012	-0.980 ***	-0.094
	(0.189)	(0.146)	(0.119)	(0.158)	(0.257)	(0.155)	(0.176)	(0.190)	(0.147)
Sec Sch Tech	0.284 **	-0.075	-0.124 *	0.371 ***	-0.005	0.182 *	0.425 ***	0.414 ***	0.145
	(0.122)	(0.093)	(0.067)	(0.091)	(0.213)	(0.101)	(0.131)	(0.130)	(0.113)
Other Major	0.566 ***	0.844 ***	0.381 ***	0.827 ***	0.477 ***	0.717 ***	0.854 ***	0.110	0.614 ***
	(0.160)	(0.104)	(0.083)	(0.121)	(0.162)	(0.093)	(0.124)	(0.107)	(0.141)
Reg Fee	-0.845 ***	-0.674 ***	-0.268 *	-0.511 ***	-0.575 *	-0.913 ***	-0.771 ***	-0.214	-0.556 ***
	(0.150)	(0.138)	(0.158)	(0.154)	(0.300)	(0.187)	(0.124)	(0.244)	(0.120)

TABLE 6. Coefficients of entry equations (Model I, cont.)

	Technologies	Exact Sciences	Engineering and Arch.	Natural and Earth Sc.	Arts	Social Sciences	Humanities	Health and Biol. Sc.	Other Health and Biol. Sc.
Prep Course	0.187 *	0.194 ***	0.200 ***	0.242 ***	0.227 *	0.144 **	0.122	0.012	0.294 ***
	(0.097)	(0.070)	(0.044)	(0.072)	(0.119)	(0.065)	(0.089)	(0.076)	(0.074)
ENEM	0.024 ***	0.038 ***	0.084 ***	0.088 ***	0.034 ***	0.057 ***	0.039 ***	0.133 ***	0.069 ***
	(0.003)	(0.006)	(0.009)	(0.007)	(0.005)	(0.007)	(0.004)	(0.008)	(0.005)
ENEM * Gender	-0.010 **	-0.020 ***	-0.025 **	-0.047 ***	-0.011 *	-0.009	-0.023 ***	-0.013	-0.028 ***
	(0.004)	(0.006)	(0.010)	(0.008)	(0.006)	(0.008)	(0.004)	(0.013)	(0.007)
Constant	-1.686 ***	-4.384 ***	-10.316 ***	-9.908 ***	-4.260 ***	-7.189 ***	-4.470 ***	-17.482 ***	-7.208 ***
	(0.389)	(0.530)	(0.944)	(0.770)	(0.540)	(0.686)	(0.405)	(0.897)	(0.508)
Number of obs.	120,058								

Estimations include the explanatory variables listed in Table 4. * significant at 10%, ** significant at 5%, *** significant at 1%.

TABLE 7. Marginal effects on the probability of entry (Model I)

	Technologies	Exact Sciences	Engineering and Arch.	Natural and Earth Sc.	Arts	Social Sciences	Humanities	Health and Biol. Sc.	Other Health and Biol. Sc.
Gender	-4.78 **	-2.04	-0.75 *	0.53	-2.75 *	-1.75 ***	-1.81	-0.71 ***	-7.80 ***
	(2.20)	(1.29)	(0.40)	(0.84)	(1.41)	(0.53)	(1.80)	(0.18)	(1.20)
White	2.58	1.36	0.36	-0.67	3.73 **	-0.12	1.97	0.49 ***	1.22
	(2.09)	(1.20)	(0.36)	(0.90)	(1.54)	(0.61)	(1.62)	(0.17)	(1.11)
Work	-0.33	-4.63 ***	-2.53 ***	-2.59 **	4.61 ***	-0.87	-1.22	0.06	-5.00 ***
	(2.27)	(1.28)	(0.49)	(1.08)	(1.56)	(0.74)	(1.63)	(0.36)	(1.28)
Age1	-12.73 ***	-15.30 ***	-2.54	-0.12	-15.98 ***	-5.20 **	-14.39 ***	1.71 ***	-11.35 ***
	(4.68)	(3.00)	(1.59)	(2.43)	(4.27)	(2.11)	(3.36)	(0.40)	(3.13)
Age2	-7.37 **	-14.25 ***	-2.34	-0.50	-14.70 ***	-4.25 **	-12.10 ***	1.64 ***	-8.49 ***
	(3.66)	(2.67)	(1.53)	(2.13)	(4.04)	(1.95)	(2.78)	(0.30)	(2.79)
Age3	-1.78	-10.99 ***	-3.00 **	-0.82	-10.54 ***	-4.04 **	-7.65 ***	0.94 ***	-4.93 *
	(3.49)	(2.63)	(1.51)	(2.11)	(3.99)	(1.94)	(2.75)	(0.28)	(2.73)
Prim Sch Priv	-1.12	4.72 ***	3.52 ***	1.26	0.58	2.36 ***	4.90 **	0.35	3.10 **
	(3.03)	(1.59)	(0.51)	(1.18)	(2.00)	(0.80)	(2.23)	(0.27)	(1.34)
Prim Sch Mixed	1.21	1.30	1.78 ***	0.66	1.52	0.64	1.77	-0.04	1.49
	(3.08)	(1.59)	(0.54)	(1.25)	(2.05)	(0.84)	(2.22)	(0.28)	(1.43)
Sec Sch Priv	3.43	4.19 ***	-4.62 ***	-5.08 ***	-0.09	-4.89 ***	2.62	-2.08 ***	1.40
	(2.87)	(1.43)	(0.58)	(1.19)	(1.89)	(0.90)	(2.14)	(0.37)	(1.35)
Sec Sch Mixed	-3.32	1.80	-4.62 ***	-4.81 **	-3.28	-7.01 ***	0.20	-2.81 ***	-1.32
	(3.90)	(2.26)	(0.94)	(1.91)	(2.60)	(1.23)	(2.97)	(0.46)	(2.06)
Sec Sch Tech	6.00 **	-1.20	-0.94 *	4.80 ***	-0.06	1.71 *	7.80 ***	1.29 ***	2.16
	(2.58)	(1.46)	(0.50)	(1.25)	(2.44)	(0.99)	(2.51)	(0.46)	(1.71)

TABLE 7. Marginal effects on the probability of entry (Model I, cont.)

	Technologies	Exact Sciences	Engineering and Arch.	Natural and Earth Sc.	Arts	Social Sciences	Humanities	Health and Biol. Sc.	Other Health and Biol. Sc.
Other Major	11.95 ***	15.26 ***	3.30 ***	11.55 ***	6.13 ***	7.71 ***	16.40 ***	0.31	9.71 ***
	(3.33)	(2.04)	(0.81)	(1.92)	(2.29)	(1.19)	(2.51)	(0.31)	(2.38)
Reg Fee	-17.11 ***	-9.56 ***	-1.95 *	-5.59 ***	-5.68 **	-6.24 ***	-11.84 ***	-0.54	-7.49 ***
	(2.78)	(1.69)	(1.06)	(1.50)	(2.49)	(0.93)	(1.66)	(0.56)	(1.47)
Prep Course	3.95 *	3.11 ***	1.54 ***	2.91 ***	2.57 *	1.27 **	2.12	0.03	4.28 ***
	(2.06)	(1.12)	(0.33)	(0.85)	(1.34)	(0.57)	(1.55)	(0.20)	(1.06)
ENEM	0.40 ***	0.44 ***	0.55 ***	0.74 ***	0.33 ***	0.47 ***	0.47 ***	0.34 ***	0.79 ***
	(0.04)	(0.04)	(0.04)	(0.06)	(0.04)	(0.04)	(0.04)	(0.02)	(0.06)
Number of obs.	120,058								

* significant at 10%, ** significant at 5%, *** significant at 1%.

TABLE 8. Marginal effects on the probability of entry for different ENEM groups (Model I)

	Technologies	Exact Sciences	Engineering and Arch.	Natural and Earth Sc.	Arts	Social Sciences	Humanities	Health and Biol. Sc.	Other Health and Biol. Sc.
= 0	7.67 **	6.67 ***	0.05 *	0.78 *	1.58 *	0.11	10.18 ***	0.00	0.84 *
	(3.40)	(1.19)	(0.03)	(0.41)	(0.92)	(0.14)	(1.67)	(0.00)	(0.48)
(0, 70]	-0.07	4.50 ***	0.94 **	5.50 ***	0.60	0.33	6.30 ***	0.02	2.15 *
	(1.96)	(1.38)	(0.37)	(1.04)	(0.90)	(0.70)	(1.36)	(0.03)	(1.12)
(70, 80]	-3.91 *	2.14	1.68 **	7.78 ***	-0.61	-0.05	2.91 *	0.03	-0.52
	(2.18)	(1.38)	(0.73)	(1.17)	(1.09)	(0.96)	(1.57)	(0.14)	(1.18)
(80, 90]	-6.33 **	-1.13	1.44 *	6.10 ***	-2.17	-0.93	-1.27	-0.04	-5.59 ***
	(2.63)	(1.38)	(0.84)	(1.01)	(1.37)	(0.84)	(1.93)	(0.29)	(1.13)
(90, 100]	-8.47 ***	-5.25 ***	-0.30	0.66	-4.24 **	-2.41 ***	-6.18 **	-0.49	-12.27 ***
	(3.15)	(1.95)	(0.63)	(1.21)	(1.99)	(0.72)	(2.58)	(0.41)	(1.90)
(100, 110]	-10.31 ***	-9.98 ***	-4.08 ***	-8.22 ***	-6.81 **	-4.45 ***	-11.46 ***	-1.88 ***	-19.29 ***
	(3.65)	(2.99)	(1.26)	(2.57)	(2.96)	(1.52)	(3.37)	(0.38)	(3.17)
(110, 120]	-11.53 ***	-14.36 ***	-9.11 ***	-17.78 ***	-9.46 **	-6.65 **	-16.23 ***	-4.42 ***	-24.77 ***
	(3.97)	(4.00)	(3.01)	(4.17)	(4.05)	(2.84)	(4.10)	(1.23)	(4.34)
Average	-4.78 **	-2.04	-0.75 *	0.53	-2.75 *	-1.75 ***	-1.81	-0.71 ***	-7.80 ***
	(2.20)	(1.29)	(0.40)	(0.84)	(1.41)	(0.53)	(1.80)	(0.18)	(1.20)
Number of obs.	120,058								

* significant at 10%, ** significant at 5%, *** significant at 1%.

TABLE 9. Coefficients of choice equations (Model I)

	Technologies	Exact Sciences	Engineering and Arch.	Natural and Earth Sc.	Arts	Social Sciences	Humanities	Health and Biol. Sc.	Other Health and Biol. Sc.
Gender	1.871 ***	6.802 ***	10.601 ***	-0.599	-2.394 ***	2.777 ***	-4.148 ***	2.325 *	-3.212 ***
	(0.329)	(0.486)	(0.657)	(0.401)	(0.800)	(0.554)	(0.451)	(1.357)	(0.409)
White	0.046	0.010	-0.801 ***	0.006	3.606 ***	0.302	0.347 **	-0.149	0.753 ***
	(0.126)	(0.157)	(0.230)	(0.172)	(0.450)	(0.222)	(0.169)	(0.594)	(0.139)
Work	1.648 ***	1.923 ***	1.844 **	2.053 ***	7.649 ***	3.770 ***	3.145 ***	0.612	1.400 ***
	(0.227)	(0.407)	(0.884)	(0.379)	(0.444)	(0.469)	(0.239)	(1.155)	(0.228)
Age1	-1.992 ***	0.503	8.319 ***	-0.239	-4.509 ***	0.213	-3.155 ***	-22.389 ***	-3.685 ***
	(0.291)	(0.321)	(0.642)	(0.402)	(0.725)	(0.527)	(0.362)	(1.896)	(0.330)
Age2	-0.487 **	0.636 **	7.524 ***	1.231 ***	-1.880 ***	3.310 ***	-1.343 ***	-17.289 ***	-0.785 ***
	(0.207)	(0.270)	(0.586)	(0.333)	(0.599)	(0.425)	(0.277)	(1.774)	(0.261)
Age3	-0.057	-0.318	3.304 ***	0.412	-0.267	1.402 ***	-0.880 ***	-12.706 ***	-0.064
	(0.189)	(0.253)	(0.572)	(0.318)	(0.522)	(0.403)	(0.245)	(1.719)	(0.250)
Prim Sch Priv	-0.751 ***	0.441 **	-1.834 ***	-0.625 ***	0.018	0.554 *	-0.153	-1.810 **	-0.787 ***
	(0.170)	(0.216)	(0.380)	(0.229)	(0.472)	(0.329)	(0.221)	(0.778)	(0.171)
Prim Sch Mixed	-0.415 **	0.313	-1.696 ***	-0.323	1.069 **	-0.094	-0.264	-0.769	-0.482 ***
	(0.168)	(0.230)	(0.403)	(0.239)	(0.477)	(0.355)	(0.230)	(0.822)	(0.182)
Sec Sch Priv	-1.312 ***	-0.783 **	-2.832 ***	-3.133 ***	-3.360 ***	-4.659 ***	-1.235 ***	-2.096 **	-0.837 ***
	(0.225)	(0.399)	(0.509)	(0.295)	(0.535)	(0.394)	(0.258)	(0.861)	(0.199)
Sec Sch Mixed	-1.009 **	0.203	-2.171 *	-1.103 *	-3.059 ***	-1.987 **	-0.343	6.056 *	-0.385
	(0.439)	(0.652)	(1.270)	(0.617)	(0.967)	(0.951)	(0.424)	(3.199)	(0.363)
Sec Sch Tech	1.623 ***	2.404 ***	3.972 ***	3.037 ***	-1.108	0.658	0.689 **	0.040	0.687 ***
	(0.257)	(0.460)	(0.737)	(0.348)	(0.820)	(0.553)	(0.316)	(1.121)	(0.262)
Other Major	0.377	4.452 ***	3.931 ***	2.532 ***	3.158 ***	5.002 ***	1.955 ***	-0.588	0.613 ***
	(0.358)	(0.345)	(0.561)	(0.325)	(0.509)	(0.378)	(0.274)	(1.244)	(0.215)
Reg Fee	-1.619 ***	1.232	-3.983 *	0.769	-9.486 ***	2.920 **	7.215 ***	9.969 ***	3.994 ***
	(0.369)	(0.863)	(2.126)	(0.797)	(2.359)	(1.411)	(0.512)	(2.047)	(0.427)
Prof F. Non-manual	0.437 **	0.659 ***	0.405	0.697 ***	0.450	0.324	0.259	-0.320	0.381 **
	(0.172)	(0.177)	(0.276)	(0.208)	(0.363)	(0.268)	(0.203)	(0.646)	(0.153)
Prof Father Manual	0.583 ***	0.804 ***	1.212 **	1.486 ***	-0.702	0.058	0.358	-1.233	0.874 ***
	(0.219)	(0.292)	(0.513)	(0.307)	(0.740)	(0.493)	(0.303)	(1.126)	(0.237)
Prof Father Other	0.641 ***	0.754 ***	1.456 ***	1.498 ***	-2.887 ***	1.161 ***	0.264	1.127	0.917 ***
	(0.205)	(0.264)	(0.481)	(0.298)	(0.759)	(0.407)	(0.282)	(1.043)	(0.211)

TABLE 9. Coefficients of choice equations (Model I, cont.)

	Technologies	Exact Sciences	Engineering and Arch.	Natural and Earth Sc.	Arts	Social Sciences	Humanities	Health and Biol. Sc.	Other Health and Biol. Sc.
Prof Mother Non-manual	0.425 **	1.326 ***	1.048 ***	0.696 ***	0.440	0.901 ***	-0.363 *	1.087 *	0.599 ***
	(0.182)	(0.194)	(0.264)	(0.213)	(0.357)	(0.271)	(0.212)	(0.645)	(0.155)
Prof Mother Manual	-0.303	1.700 ***	1.316 **	1.152 ***	2.246 ***	1.918 ***	0.218	3.285 *	0.505
	(0.336)	(0.397)	(0.633)	(0.393)	(0.777)	(0.549)	(0.368)	(1.786)	(0.324)
Prof Mother Housewife	0.562 ***	1.647 ***	1.422 ***	0.831 ***	0.229	0.662 **	-0.801 ***	-0.191	0.491 ***
	(0.186)	(0.197)	(0.273)	(0.224)	(0.378)	(0.274)	(0.223)	(0.676)	(0.165)
Prof Mother Other	0.244	1.758 ***	1.456 ***	0.409	0.378	1.053 ***	0.499 *	1.552	0.396 *
	(0.228)	(0.264)	(0.456)	(0.327)	(0.605)	(0.406)	(0.277)	(1.122)	(0.234)
Rsn Maj Job Mkt	3.273 ***	3.357 ***	2.712 ***	3.979 ***	-30.411 **	1.768 ***	-2.472 ***	0.393	-0.236
	(0.269)	(0.411)	(0.658)	(0.415)	(13.329)	(0.630)	(0.577)	(1.486)	(0.349)
Rsn Maj Soc Cont	-0.382	-9.485 ***	-19.753 ***	-7.390 ***	-9.369 ***	-3.449 ***	-1.716 ***	-3.478 ***	-1.464 ***
	(0.314)	(1.242)	(1.788)	(0.697)	(1.269)	(0.447)	(0.281)	(1.002)	(0.213)
Rsn Maj Pers Real	-1.040 ***	-2.048 ***	-3.687 ***	-1.537 ***	-1.261 ***	-2.368 ***	-1.112 ***	-3.638 ***	-0.294 **
	(0.293)	(0.331)	(0.391)	(0.268)	(0.364)	(0.320)	(0.213)	(0.692)	(0.148)
Rsn Maj Other	3.122 ***	0.449	-0.249	2.397 ***	-5.325 ***	0.889	0.949 ***	-1.453	1.337 ***
	(0.256)	(0.525)	(0.703)	(0.414)	(1.513)	(0.565)	(0.321)	(1.345)	(0.256)
Rsn Univ Free	0.789 ***	-0.435	0.872 *	-3.002 ***	-3.930 ***	0.756 *	0.790 ***	2.739 ***	0.283
	(0.256)	(0.405)	(0.492)	(0.336)	(0.537)	(0.388)	(0.238)	(0.891)	(0.178)
Rsn Univ Rep	0.924 ***	0.483	2.188 ***	-2.254 ***	-3.919 ***	1.172 ***	0.599 ***	2.703 ***	0.035
	(0.253)	(0.336)	(0.401)	(0.281)	(0.457)	(0.335)	(0.224)	(0.764)	(0.159)
Rsn Univ Other	1.198 ***	1.912 ***	1.404 ***	-0.874 ***	-2.191 ***	1.157 ***	0.863 ***	1.453	0.076
	(0.265)	(0.341)	(0.494)	(0.302)	(0.485)	(0.396)	(0.252)	(0.930)	(0.189)
Work * Gender	-0.329	-0.441	1.031	-1.430 ***	1.622 **	-2.097 ***	-0.891 **	-6.255 ***	-2.070 ***
	(0.273)	(0.456)	(0.987)	(0.458)	(0.661)	(0.586)	(0.369)	(1.609)	(0.444)
Sec Sch Priv * Gender	0.991 ***	-0.892 **	2.835 ***	1.090 ***	1.514 **	2.681 ***	1.281 ***	7.456 ***	1.146 ***
	(0.264)	(0.440)	(0.591)	(0.363)	(0.726)	(0.487)	(0.380)	(1.146)	(0.369)
Sec Sch Mixed * Gender	0.635	-1.089	4.153 ***	-0.259	0.824	1.175	1.337 **	3.798	1.022
	(0.528)	(0.746)	(1.450)	(0.783)	(1.441)	(1.246)	(0.634)	(3.896)	(0.664)

TABLE 9. Coefficients of choice equations (Model I, cont.)

	Technologies	Exact Sciences	Engineering and Arch.	Natural and Earth Sc.	Arts	Social Sciences	Humanities	Health and Biol. Sc.	Other Health and Biol. Sc.
Sec Sch Tech * Gender	-0.501	-0.611	-0.183	-1.673 ***	-1.428	-0.282	0.163	-5.417 ***	-0.428
	(0.309)	(0.527)	(0.900)	(0.441)	(1.114)	(0.696)	(0.452)	(1.669)	(0.499)
Other Major * Gender	0.614	-2.127 ***	-3.804 ***	-0.014	-0.947	-2.193 ***	2.431 ***	-2.748 *	1.275 ***
	(0.393)	(0.386)	(0.654)	(0.392)	(0.712)	(0.474)	(0.370)	(1.625)	(0.369)
Reg Fee * Gender	1.673 ***	4.757 ***	11.622 ***	1.131	7.464 **	3.321 *	0.407	-4.130	1.226
	(0.497)	(1.022)	(2.371)	(0.983)	(3.101)	(1.876)	(0.728)	(2.994)	(0.810)
Rsn Maj J Mkt * Gender	-1.134 ***	-2.432 ***	-0.061	-1.380 ***	-10.954	-0.378	-1.832 *	1.510	-2.236 ***
	(0.325)	(0.495)	(0.822)	(0.529)	(15.883)	(0.816)	(1.108)	(2.028)	(0.782)
Rsn Maj S Cont * Gender	-1.323 ***	4.506 ***	7.607 ***	5.011 ***	-2.598	1.873 ***	1.824 ***	1.825	-0.488
	(0.432)	(1.299)	(1.873)	(0.787)	(2.377)	(0.633)	(0.440)	(1.521)	(0.491)
Rsn Maj P Real * Gender	-0.225	1.556 ***	1.228 **	1.172 ***	1.824 ***	1.465 ***	0.631 *	2.735 ***	0.586 *
	(0.366)	(0.382)	(0.496)	(0.365)	(0.580)	(0.456)	(0.378)	(1.054)	(0.318)
Rsn Maj Other * Gender	-1.340 ***	-0.598	-1.148	-1.511 ***	1.747	-1.174	-0.290	2.994	-0.209
	(0.324)	(0.603)	(0.848)	(0.549)	(1.954)	(0.757)	(0.528)	(1.836)	(0.479)
Rsn Univ Free * Gender	-0.576 *	-1.097 **	-2.564 ***	1.953 ***	-0.400	-0.161	-0.332	-0.102	0.626 *
	(0.310)	(0.468)	(0.607)	(0.442)	(0.827)	(0.532)	(0.426)	(1.300)	(0.353)
Rsn Univ Rep * Gender	-1.237 ***	-1.567 ***	-3.000 ***	1.331 ***	-3.249 ***	-0.215	0.030	-1.530	-0.275
	(0.315)	(0.392)	(0.502)	(0.379)	(0.866)	(0.462)	(0.399)	(1.116)	(0.341)
Rsn Univ Other * Gender	-0.942 ***	-1.177 ***	-1.956 ***	1.842 ***	-0.854	-0.687	0.766 *	-1.039	-0.559
	(0.325)	(0.397)	(0.612)	(0.398)	(0.763)	(0.552)	(0.406)	(1.349)	(0.403)
Constant	-5.743 ***	-11.203 ***	-9.306 ***	-4.425 ***	-7.956 ***	-8.838 ***	-6.131 ***	19.486 ***	-3.539 ***
	(0.408)	(0.568)	(0.892)	(0.531)	(0.980)	(0.712)	(0.473)	(2.265)	(0.395)
Number of observations	120,058								

Estimations include the explanatory variables listed in Table 4. * significant at 10%, ** significant at 5%, *** significant at 1%.

TABLE 10. Marginal effects on choice probabilities (Model I)

	Technologies	Exact Sciences	Engineering and Arch.	Natural and Earth Sc.	Arts	Social Sciences	Humanities	Health and Biol. Sc.	Other Health and Biol. Sc.
Gender	0.73 ***	4.10 ***	13.86 ***	-2.82 ***	-0.78 ***	1.69 ***	-3.04 ***	-10.38 ***	-3.35 ***
	(0.08)	(0.13)	(0.19)	(0.14)	(0.10)	(0.17)	(0.10)	(0.21)	(0.12)
White	-0.20 **	-0.21	-1.42 ***	0.29 *	-0.13	0.61 ***	-0.06	0.49 **	0.63 ***
	(0.09)	(0.14)	(0.23)	(0.15)	(0.11)	(0.18)	(0.11)	(0.24)	(0.12)
Work	0.04	-1.06 ***	16.13 ***	-2.54 ***	0.87 ***	-2.16 ***	1.09 ***	-11.39 ***	-0.97 ***
	(0.09)	(0.17)	(0.51)	(0.18)	(0.21)	(0.29)	(0.16)	(0.40)	(0.17)
Prim Sch Priv	-0.36 ***	-0.18	-0.21	-0.36 *	0.07	0.58 **	-0.62 ***	2.28 ***	-1.20 ***
	(0.11)	(0.19)	(0.33)	(0.19)	(0.14)	(0.26)	(0.16)	(0.33)	(0.18)
Prim Sch Mixed	-0.36 ***	0.32	-0.97 ***	-0.08	0.07	0.32	-0.32 **	1.65 ***	-0.63 ***
	(0.12)	(0.20)	(0.35)	(0.20)	(0.15)	(0.27)	(0.16)	(0.35)	(0.19)
Sec Sch Priv	-3.57 ***	3.80 ***	11.17 ***	-2.14 ***	-1.42 ***	-7.60 ***	-0.46	0.51 ***	-0.30
	(0.49)	(1.48)	(1.04)	(0.26)	(0.28)	(0.80)	(0.33)	(0.14)	(0.21)
Sec Sch Mixed	-4.44 ***	-8.26 ***	13.08 ***	-1.57 ***	-1.07 ***	-5.56 ***	-0.36	8.27 ***	-0.09
	(0.82)	(1.29)	(1.43)	(0.36)	(0.39)	(1.02)	(0.46)	(0.66)	(0.30)
Sec Sch Tech	-0.34 ***	-0.98 ***	21.11 ***	-1.75 ***	-1.34 ***	-4.85 ***	-1.10 ***	-8.93 ***	-1.80 ***
	(0.10)	(0.21)	(0.49)	(0.22)	(0.12)	(0.27)	(0.14)	(0.45)	(0.19)
Other Major	-1.11 ***	-1.04 ***	14.59 ***	-2.33 ***	-1.23 ***	-0.86 **	-0.16	-5.41 ***	-2.45 ***
	(0.09)	(0.25)	(0.62)	(0.27)	(0.13)	(0.38)	(0.18)	(0.50)	(0.17)
Reg Fee	-0.35 ***	0.31	15.08 ***	-3.73 ***	-1.47 ***	-4.43 ***	7.24 ***	-14.11 ***	1.47 ***
	(0.09)	(0.46)	(0.89)	(0.24)	(0.16)	(0.39)	(0.53)	(0.64)	(0.39)
Prof Father Non-manual	0.24 **	0.62 ***	0.01	0.39 **	0.21 *	-0.37	0.15	-1.17 ***	-0.08
	(0.11)	(0.17)	(0.27)	(0.18)	(0.13)	(0.23)	(0.14)	(0.28)	(0.15)
Prof Father Manual	0.33 **	0.87 ***	1.00 *	0.29	0.30	0.03	-0.09	-3.03 ***	0.29
	(0.15)	(0.28)	(0.53)	(0.29)	(0.23)	(0.40)	(0.20)	(0.52)	(0.25)
Prof Father Other	0.30 **	0.94 ***	-0.33	0.69 **	0.52 ***	0.50	0.26	-2.73 ***	-0.15
	(0.14)	(0.24)	(0.43)	(0.27)	(0.20)	(0.34)	(0.18)	(0.43)	(0.21)

TABLE 10. Marginal effects on choice probabilities (Model I, cont.)

	Technologies	Exact Sciences	Engineering and Arch.	Natural and Earth Sc.	Arts	Social Sciences	Humanities	Health and Biol. Sc.	Other Health and Biol. Sc.
Prof Mother Non-manual	0.11	0.26	0.39	0.61 ***	0.20 *	0.06	-0.09	-1.78 ***	0.24
	(0.12)	(0.16)	(0.26)	(0.19)	(0.12)	(0.21)	(0.15)	(0.27)	(0.15)
Prof Mother Manual	0.14	0.50	0.56	0.38	-0.26	0.17	0.08	-1.48 **	-0.09
	(0.20)	(0.37)	(0.66)	(0.36)	(0.30)	(0.58)	(0.26)	(0.66)	(0.32)
Prof Mother Housewife	0.15	0.46 ***	1.72 ***	0.95 ***	-0.47 ***	-1.10 ***	-0.36 **	-1.88 ***	0.54 ***
	(0.13)	(0.17)	(0.27)	(0.20)	(0.12)	(0.23)	(0.15)	(0.28)	(0.16)
Prof Mother Other	0.17	0.12	0.90 **	0.67 **	0.20	-0.32	-0.03	-1.79 ***	0.08
	(0.15)	(0.26)	(0.43)	(0.27)	(0.20)	(0.36)	(0.20)	(0.44)	(0.22)
Rsn Maj Job Mkt	2.55 ***	1.52 ***	2.43 ***	3.71 ***	-3.32 ***	0.14	-1.71 ***	-4.02 ***	-1.29 ***
	(0.20)	(0.28)	(0.43)	(0.32)	(0.21)	(0.35)	(0.16)	(0.42)	(0.23)
Rsn Maj Soc Cont	-0.17	-2.87 ***	-12.07 ***	-1.06 ***	-1.73 ***	3.12 ***	0.48 ***	13.84 ***	0.45 **
	(0.12)	(0.19)	(0.33)	(0.23)	(0.16)	(0.30)	(0.16)	(0.39)	(0.19)
Rsn Maj Pers Real	-0.49 ***	-0.29 *	-2.97 ***	0.05	0.23 **	-0.14	-0.26 **	3.14 ***	0.72 ***
	(0.08)	(0.15)	(0.23)	(0.16)	(0.11)	(0.19)	(0.11)	(0.24)	(0.14)
Rsn Maj Other	2.36 ***	-0.18	-3.17 ***	1.72 ***	-1.44 ***	-0.10	0.44 **	-1.07 **	1.43 ***
	(0.20)	(0.25)	(0.40)	(0.32)	(0.19)	(0.34)	(0.20)	(0.43)	(0.27)
Rsn Univ Free	0.37 ***	-1.05 ***	-1.54 ***	-2.02 ***	-1.47 ***	1.29 ***	0.54 ***	3.12 ***	0.75 ***
	(0.10)	(0.16)	(0.27)	(0.18)	(0.12)	(0.23)	(0.13)	(0.29)	(0.16)
Rsn Univ Rep	0.14	-0.56 ***	0.28	-1.76 ***	-1.76 ***	1.60 ***	0.44 ***	1.66 ***	-0.04
	(0.10)	(0.15)	(0.23)	(0.16)	(0.11)	(0.19)	(0.12)	(0.24)	(0.14)
Rsn Univ Other	0.38 ***	1.16 ***	-0.86 ***	-0.32 *	-1.18 ***	0.61 ***	0.70 ***	-0.15	-0.35 **
	(0.11)	(0.18)	(0.28)	(0.19)	(0.13)	(0.22)	(0.13)	(0.29)	(0.16)
Number of observations	120,058								

* significant at 10%, ** significant at 5%, *** significant at 1%.

TABLE 11. Coefficients of entry equations (Model II)

	Technologies	Exact Sciences	Engineering and Arch.	Natural and Earth Sc.	Arts	Social Sciences	Humanities	Health and Biol. Sc.	Other Health and Biol. Sc.
Gender	-0.538 **	-0.676 ***	5.554 ***	3.774 ***	-0.382 *	1.464 ***	0.395 **	-5.053 ***	0.218
	(0.236)	(0.216)	(0.340)	(0.365)	(0.223)	(0.318)	(0.178)	(0.442)	(0.229)
White	0.174	0.131 *	0.084 *	-0.008	0.192	-0.069	-0.030	0.050	0.143 **
	(0.124)	(0.074)	(0.044)	(0.073)	(0.118)	(0.069)	(0.061)	(0.048)	(0.066)
Work	0.027	0.034	-0.161 **	-0.155	0.164	-0.155	0.002	-0.127 *	-0.047
	(0.142)	(0.090)	(0.064)	(0.104)	(0.122)	(0.099)	(0.074)	(0.077)	(0.089)
Age1	0.083	-0.274 *	-0.129	-0.428 **	-0.145	-0.460 **	-0.688 ***	0.072	-0.550 ***
	(0.299)	(0.166)	(0.126)	(0.188)	(0.228)	(0.179)	(0.166)	(0.137)	(0.163)
Age2	-0.014	-0.321 **	-0.039	-0.281 *	-0.242	-0.347 **	-0.458 ***	0.101	-0.437 ***
	(0.225)	(0.131)	(0.115)	(0.151)	(0.187)	(0.141)	(0.096)	(0.119)	(0.130)
Age3	0.241	-0.117	0.032	-0.277 *	-0.172	-0.274 **	-0.323 ***	0.201 *	-0.295 **
	(0.219)	(0.125)	(0.115)	(0.147)	(0.181)	(0.136)	(0.091)	(0.117)	(0.126)
Prim Sch Priv	0.128	0.083	0.095	0.121	0.201	0.046	0.260 *	0.082	0.131
	(0.181)	(0.104)	(0.065)	(0.102)	(0.143)	(0.108)	(0.144)	(0.067)	(0.097)
Prim Sch Mixed	0.039	0.014	0.105	0.021	0.082	0.001	0.165	0.147 **	-0.034
	(0.196)	(0.107)	(0.070)	(0.108)	(0.138)	(0.114)	(0.124)	(0.071)	(0.107)
Sec Sch Priv	0.101	0.259 **	-0.172 ***	-0.165	0.180	-0.046	0.187	-0.134 *	0.296 **
	(0.177)	(0.112)	(0.061)	(0.105)	(0.132)	(0.102)	(0.154)	(0.068)	(0.118)
Sec Sch Mixed	-0.325	-0.038	-0.152	-0.207	-0.083	-0.570 **	-0.039	-0.424 ***	-0.305
	(0.306)	(0.181)	(0.109)	(0.190)	(0.203)	(0.232)	(0.234)	(0.134)	(0.210)
Sec Sch Tech	-0.184	-0.048	0.094	0.361 ***	-0.486 **	-0.085	-0.041	-0.186 **	0.000
	(0.176)	(0.113)	(0.069)	(0.101)	(0.238)	(0.121)	(0.099)	(0.094)	(0.125)
Other Major	0.598 ***	0.309 ***	0.198 ***	0.398 ***	-0.074	0.358 ***	0.205 **	0.342 ***	0.665 ***
	(0.219)	(0.098)	(0.072)	(0.112)	(0.148)	(0.097)	(0.096)	(0.072)	(0.107)
Reg Fee	-1.077 ***	-0.841 ***	-0.473 ***	-0.827 ***	-0.168	-1.727 ***	-0.591 ***	-0.233 *	-0.429 **
	(0.210)	(0.184)	(0.137)	(0.213)	(0.202)	(0.441)	(0.167)	(0.139)	(0.175)

TABLE 11. Coefficients of entry equations (Model II, cont.)

	Technologies	Exact Sciences	Engineering and Arch.	Natural and Earth Sc.	Arts	Social Sciences	Humanities	Health and Biol. Sc.	Other Health and Biol. Sc.
Prep Course	0.259 **	0.017	-0.007	0.196 ***	0.371 ***	0.286 ***	0.106 *	0.301 ***	0.265 ***
	(0.101)	(0.066)	(0.024)	(0.067)	(0.078)	(0.057)	(0.060)	(0.025)	(0.062)
ENEM	0.006 ***	-0.000	0.059 ***	0.048 ***	0.002	0.041 ***	0.005 ***	0.021 ***	0.016 ***
	(0.002)	(0.002)	(0.003)	(0.003)	(0.002)	(0.002)	(0.001)	(0.001)	(0.002)
ENEM * Gender	0.003	0.010 ***	-0.052 ***	-0.037 ***	0.005 **	-0.016 ***	-0.002	0.044 ***	-0.005 *
	(0.003)	(0.002)	(0.003)	(0.004)	(0.002)	(0.003)	(0.002)	(0.004)	(0.002)
Constant	-0.818 **	-1.269 ***	-8.368 ***	-6.287 ***	-2.371 ***	-5.722 ***	-1.657 ***	-5.615 ***	-3.063 ***
	(0.365)	(0.275)	(0.359)	(0.390)	(0.327)	(0.308)	(0.222)	(0.187)	(0.250)
Number of obs.	39,494								

Estimations include the explanatory variables listed in Table 4. * significant at 10%, ** significant at 5%, *** significant at 1%.

TABLE 12. Marginal effects on the probability of entry (Model II)

	Technologies	Exact Sciences	Engineering and Arch.	Natural and Earth Sc.	Arts	Social Sciences	Humanities	Health and Biol. Sc.	Other Health and Biol. Sc.
Gender	-7.08 *	3.21	2.93 ***	2.05 *	1.31	-1.46 **	4.43	-1.47 ***	-3.37 **
	(3.73)	(2.08)	(0.47)	(1.20)	(1.70)	(0.74)	(2.69)	(0.19)	(1.64)
White	4.01	2.27 *	0.62 *	-0.10	2.18 *	-0.59	-0.55	0.14	2.20 **
	(2.85)	(1.25)	(0.33)	(0.96)	(1.29)	(0.60)	(1.14)	(0.13)	(1.01)
Work	0.61	0.61	-1.17 ***	-1.98	2.01	-1.26	0.04	-0.34 *	-0.73
	(3.27)	(1.60)	(0.44)	(1.29)	(1.55)	(0.77)	(1.38)	(0.19)	(1.38)
Age1	1.92	-5.07	-0.96	-5.88 **	-1.86	-4.20 **	-13.15 ***	0.19	-9.18 ***
	(6.90)	(3.08)	(0.97)	(2.68)	(2.96)	(1.71)	(3.08)	(0.35)	(2.81)
Age2	-0.32	-5.86 **	-0.30	-4.02 *	-3.02	-3.29 **	-9.13 ***	0.26	-7.48 ***
	(5.18)	(2.51)	(0.90)	(2.29)	(2.47)	(1.47)	(2.02)	(0.30)	(2.41)
Age3	5.53	-2.23	0.26	-3.97 *	-2.19	-2.66 *	-6.59 ***	0.55 *	-5.21 **
	(5.03)	(2.41)	(0.90)	(2.22)	(2.39)	(1.41)	(1.91)	(0.30)	(2.33)
Prim Sch Priv	2.96	1.46	0.71	1.58	2.31	0.39	4.76 *	0.22	2.06
	(4.20)	(1.82)	(0.47)	(1.32)	(1.62)	(0.91)	(2.60)	(0.18)	(1.51)
Prim Sch Mixed	0.90	0.24	0.78	0.27	0.91	0.01	2.95	0.41 **	-0.52
	(4.53)	(1.84)	(0.52)	(1.37)	(1.52)	(0.95)	(2.24)	(0.20)	(1.60)
Sec Sch Priv	2.34	4.46 **	-1.35 ***	-2.24	2.07	-0.40	3.43	-0.39 *	4.56 **
	(4.11)	(1.90)	(0.50)	(1.45)	(1.49)	(0.90)	(2.79)	(0.21)	(1.77)
Sec Sch Mixed	-7.47	-0.62	-1.20	-2.78	-0.87	-4.11 ***	-0.68	-1.09 ***	-3.96
	(6.97)	(2.88)	(0.83)	(2.45)	(2.10)	(1.45)	(4.06)	(0.31)	(2.58)
Sec Sch Tech	-4.23	-0.84	0.74	5.18 ***	-4.96 **	-0.71	-0.76	-0.48 **	0.01
	(4.04)	(1.95)	(0.56)	(1.55)	(2.09)	(0.97)	(1.81)	(0.23)	(1.97)

TABLE 12. Marginal effects on the probability of entry (Model II, cont.)

	Technologies	Exact Sciences	Engineering and Arch.	Natural and Earth Sc.	Arts	Social Sciences	Humanities	Health and Biol. Sc.	Other Health and Biol. Sc.
Other Major	13.39 ***	5.74 ***	1.61 ***	5.77 ***	-0.86	3.38 ***	3.94 **	1.09 ***	11.89 ***
	(4.64)	(1.91)	(0.62)	(1.76)	(1.67)	(1.01)	(1.90)	(0.27)	(2.15)
Reg Fee	-24.17 ***	-12.28 ***	-3.06 ***	-8.79 ***	-1.89	-8.35 ***	-9.80 ***	-0.59 *	-6.13 ***
	(4.28)	(2.18)	(0.74)	(1.76)	(2.16)	(1.02)	(2.49)	(0.32)	(2.27)
Prep Course	5.99 ***	0.30	-0.06	2.57 ***	4.26 ***	2.36 ***	1.96 *	0.80 ***	4.12 ***
	(2.32)	(1.16)	(0.18)	(0.86)	(0.94)	(0.46)	(1.11)	(0.07)	(0.96)
ENEM	0.17 ***	0.08 ***	0.21 ***	0.37 ***	0.05 ***	0.28 ***	0.08 ***	0.11 ***	0.22 ***
	(0.03)	(0.02)	(0.01)	(0.03)	(0.01)	(0.02)	(0.02)	(0.01)	(0.02)
Number of obs.	39,494								

* significant at 10%, ** significant at 5%, *** significant at 1%.

TABLE 13. Marginal effects on the probability of entry for different ENEM groups (Model II)

	Technologies	Exact Sciences	Engineering and Arch.	Natural and Earth Sc.	Arts	Social Sciences	Humanities	Health and Biol. Sc.	Other Health and Biol. Sc.
= 0	-11.63 **	-9.67 ***	5.59 ***	7.84 ***	-3.23	0.77 ***	6.18 **	-0.51 ***	1.61
	(5.13)	(3.41)	(0.37)	(1.27)	(1.99)	(0.23)	(2.83)	(0.06)	(1.71)
(0, 70]	-8.39 **	-1.39	6.66 ***	8.61 ***	-0.52	1.12 ***	4.12 **	-1.34 ***	-0.59
	(3.26)	(1.74)	(0.37)	(0.97)	(1.28)	(0.37)	(2.01)	(0.12)	(1.21)
(70, 80]	-7.69 **	0.91	6.79 ***	8.24 ***	0.35	0.96 *	4.17 *	-1.81 ***	-1.73
	(3.51)	(1.81)	(0.38)	(1.01)	(1.39)	(0.49)	(2.27)	(0.15)	(1.34)
(80, 90]	-7.18 *	2.62	6.12 ***	6.59 ***	0.99	0.38	4.24 *	-2.06 ***	-2.68 *
	(3.80)	(1.96)	(0.41)	(1.09)	(1.57)	(0.60)	(2.52)	(0.18)	(1.51)
(90, 100]	-6.63	4.49 **	4.35 ***	3.41 ***	1.71	-0.82	4.28	-2.07 ***	-3.79 **
	(4.13)	(2.16)	(0.48)	(1.27)	(1.78)	(0.77)	(2.82)	(0.21)	(1.76)
(100, 110]	-6.00	6.54 ***	0.88	-1.76	2.55	-2.91 ***	4.25	-1.58 ***	-5.02 **
	(4.48)	(2.42)	(0.65)	(1.62)	(2.08)	(1.05)	(3.14)	(0.25)	(2.10)
(110, 120]	-5.39	8.49 ***	-4.24 ***	-8.01 ***	3.41	-5.68 ***	4.16	-0.35	-6.21 **
	(4.77)	(2.69)	(0.98)	(2.14)	(2.38)	(1.46)	(3.43)	(0.35)	(2.47)
Average	-7.08 *	3.21	2.93 ***	2.05 *	1.31	-1.46 **	4.43	-1.47 ***	-3.37 **
	(3.73)	(2.08)	(0.47)	(1.20)	(1.70)	(0.74)	(2.69)	(0.19)	(1.64)
Number of obs.	39,494								

* significant at 10%, ** significant at 5%, *** significant at 1%.

TABLE 14. Coefficients of choice equations (Model II)

	Technologies	Exact Sciences	Engineering and Arch.	Natural and Earth Sc.	Arts	Social Sciences	Humanities	Health and Biol. Sc.	Other Health and Biol. Sc.
Gender	1.534 ***	5.817 ***	8.143 ***	-4.379 ***	-3.010 *	-1.242	-6.168 ***	-8.669 *	-4.047 ***
	(0.552)	(1.035)	(1.848)	(0.991)	(1.624)	(1.473)	(1.248)	(4.739)	(0.991)
White	0.089	-0.071	-2.340 ***	0.475	1.969 **	0.181	0.031	-2.433	0.563
	(0.239)	(0.333)	(0.835)	(0.388)	(0.969)	(0.580)	(0.376)	(2.525)	(0.364)
Work	1.125 ***	1.417 **	0.340	0.050	4.076 ***	1.615	2.008 ***	-16.806 ***	0.063
	(0.344)	(0.715)	(1.857)	(0.853)	(0.957)	(1.261)	(0.549)	(4.379)	(0.552)
Age1	-0.915 *	0.428	8.423 ***	-2.719 ***	-2.811 *	-1.453	-4.614 ***	-52.531 ***	-3.773 ***
	(0.551)	(0.715)	(2.116)	(0.971)	(1.496)	(1.467)	(0.988)	(9.284)	(0.826)
Age2	-0.757 *	0.061	6.862 ***	-0.660	-2.302 *	1.509	-2.631 ***	-47.722 ***	-2.051 ***
	(0.401)	(0.614)	(1.881)	(0.778)	(1.231)	(1.234)	(0.606)	(8.608)	(0.673)
Age3	-0.527	-1.291 **	1.052	-1.179	-1.402	-0.695	-2.305 ***	-39.683 ***	-1.859 ***
	(0.375)	(0.586)	(1.818)	(0.748)	(1.158)	(1.178)	(0.559)	(8.385)	(0.643)
Prim Sch Priv	-0.256	0.390	-0.696	-0.144	1.770	0.472	-0.490	0.087	-1.385 ***
	(0.345)	(0.444)	(1.175)	(0.536)	(1.083)	(0.910)	(0.590)	(3.427)	(0.448)
Prim Sch Mixed	-0.149	1.283 ***	0.704	1.170 **	3.601 ***	2.474 **	0.322	1.787	0.282
	(0.370)	(0.471)	(1.254)	(0.580)	(1.084)	(1.005)	(0.573)	(3.568)	(0.504)
Sec Sch Priv	-0.516	0.591	3.401 **	0.053	-1.433	1.140	0.171	21.745 ***	0.980 *
	(0.398)	(0.775)	(1.475)	(0.744)	(1.207)	(1.122)	(0.657)	(3.870)	(0.589)
Sec Sch Mixed	-0.797	1.430	2.359	-0.105	1.789	4.296	1.035	29.240 ***	2.293 *
	(0.854)	(1.302)	(2.796)	(1.354)	(1.901)	(3.096)	(1.014)	(9.737)	(1.274)
Sec Sch Tech	1.102 **	0.949	4.701 **	2.099 **	-4.171	0.603	0.164	4.966	-0.079
	(0.438)	(1.030)	(1.986)	(0.846)	(2.648)	(1.596)	(0.796)	(6.013)	(0.745)
Other Major	-0.617	2.213 ***	-1.349	-1.730 **	0.779	-0.687	-0.001	-25.782 ***	-2.455 ***
	(0.513)	(0.733)	(1.782)	(0.810)	(1.178)	(1.057)	(0.630)	(3.969)	(0.546)
Reg Fee	-0.843	4.718 ***	9.701 **	6.264 **	0.261	52.150 **	16.634 ***	10.506	8.621 ***
	(0.543)	(1.535)	(4.050)	(2.743)	(2.034)	(25.258)	(2.182)	(8.607)	(1.760)
Prof Father Non-manual	0.531	1.244 ***	1.016	1.081 **	0.817	0.855	0.387	-0.716	0.069
	(0.388)	(0.388)	(0.957)	(0.475)	(0.837)	(0.746)	(0.516)	(2.899)	(0.382)
Prof Father Manual	0.561	1.647 ***	3.258 *	2.686 ***	2.473 *	3.129 **	0.036	7.795	0.461
	(0.485)	(0.620)	(1.683)	(0.763)	(1.332)	(1.473)	(0.741)	(5.454)	(0.624)
Prof Father Other	1.009 **	1.851 ***	3.258 **	3.193 ***	0.662	2.483 **	1.068	10.536 **	0.755
	(0.442)	(0.568)	(1.559)	(0.742)	(1.169)	(1.093)	(0.650)	(4.782)	(0.583)

TABLE 14. Coefficients of choice equations (Model II, cont.)

	Technologies	Exact Sciences	Engineering and Arch.	Natural and Earth Sc.	Arts	Social Sciences	Humanities	Health and Biol. Sc.	Other Health and Biol. Sc.
Prof Mother Non-manual	0.378	0.321	-0.466	0.722	0.752	-0.202	0.014	-0.452	0.781 *
	(0.429)	(0.421)	(0.945)	(0.498)	(0.800)	(0.736)	(0.545)	(2.994)	(0.401)
Prof Mother Manual	0.379	-0.055	-1.695	0.623	-0.324	-2.535 *	-0.547	-8.692	-0.626
	(0.624)	(0.750)	(1.832)	(0.948)	(1.828)	(1.523)	(0.944)	(6.621)	(0.817)
Prof Mother Housewife	1.162 ***	1.569 ***	1.462	1.759 ***	0.147	0.417	-0.236	-0.234	1.342 ***
	(0.435)	(0.445)	(0.962)	(0.517)	(1.014)	(0.770)	(0.566)	(3.153)	(0.437)
Prof Mother Other	0.644	1.052 **	3.873 ***	2.077 ***	-0.089	-0.029	0.772	-3.273	1.561 ***
	(0.493)	(0.533)	(1.453)	(0.712)	(1.254)	(0.893)	(0.655)	(3.803)	(0.562)
Rsn Maj Job Mkt	2.127 ***	1.505 *	-2.622	1.913 **	-40.932 ***	-3.487 **	-5.526 ***	-20.889 ***	-3.282 ***
	(0.431)	(0.778)	(1.898)	(0.933)	(11.981)	(1.383)	(1.054)	(4.383)	(0.773)
Rsn Maj Soc Cont	1.427 ***	-6.925 ***	-6.604 ***	-0.021	-6.889 ***	4.942 ***	2.072 ***	35.109 ***	2.676 ***
	(0.463)	(1.580)	(1.953)	(1.065)	(1.780)	(1.307)	(0.711)	(5.117)	(0.662)
Rsn Maj Pers Real	0.228	-1.790 ***	-1.218	-0.020	0.070	0.207	-0.747	7.308 ***	1.046 **
	(0.392)	(0.679)	(1.135)	(0.604)	(0.797)	(0.808)	(0.492)	(2.601)	(0.411)
Rsn Maj Other	2.792 ***	-0.887	-1.464	3.271 ***	-6.297 ***	0.735	0.757	-5.197	1.412 **
	(0.432)	(0.981)	(1.986)	(0.959)	(2.044)	(1.322)	(0.737)	(4.237)	(0.660)
Rsn Univ Free	1.104 ***	-0.052	1.031	-2.771 ***	-3.590 ***	1.100	1.246 **	5.626 *	1.602 ***
	(0.379)	(0.732)	(1.430)	(0.763)	(1.099)	(0.975)	(0.582)	(3.114)	(0.485)
Rsn Univ Rep	0.677 *	0.126	1.207	-3.266 ***	-7.282 ***	-0.457	0.996 *	1.453	-0.467
	(0.380)	(0.645)	(1.197)	(0.664)	(1.216)	(0.856)	(0.523)	(2.774)	(0.454)
Rsn Univ Other	1.346 ***	1.222 *	2.633 *	-0.948	-5.194 ***	1.177	1.048 *	3.404	0.359
	(0.373)	(0.708)	(1.392)	(0.708)	(1.172)	(0.964)	(0.578)	(3.044)	(0.496)
Work * Gender	0.295	0.855	5.096 **	1.749 *	3.897 ***	2.140	1.045	13.663 ***	0.188
	(0.440)	(0.846)	(2.072)	(1.000)	(1.472)	(1.533)	(0.842)	(5.203)	(0.905)
Sec Sch Priv * Gender	-0.461	-2.763 ***	-3.511 **	-1.247	-2.490 *	-2.935 **	1.348	-12.851 ***	-1.796 *
	(0.441)	(0.913)	(1.569)	(0.886)	(1.489)	(1.298)	(0.915)	(4.150)	(0.925)
Sec Sch Mixed * Gender	-1.029	-4.082 ***	-5.279 *	-2.466	-6.299 **	-8.757 **	-0.553	-24.797 **	-6.329 **
	(0.942)	(1.563)	(3.047)	(1.771)	(3.097)	(3.933)	(1.563)	(10.451)	(2.660)

TABLE 14. Coefficients of choice equations (Model II, cont.)

	Technologies	Exact Sc. and Agr. Eng.	Engineering and Arch.	Natural and Earth Sc.	Arts	Social Sciences	Humanities	Health and Biol. Sc.	Other Health and Biol. Sc.
Sec Sch Tech * Gender	-0.216	2.461 **	1.345	-0.002	-1.265	2.371	0.933	-0.763	-0.372
	(0.598)	(1.250)	(2.236)	(1.035)	(3.447)	(2.007)	(1.097)	(6.937)	(1.294)
Other Major * Gender	1.117 *	-0.719	0.463	1.269	-0.200	2.335 *	2.683 ***	8.622 *	3.141 ***
	(0.587)	(0.827)	(1.967)	(0.972)	(1.608)	(1.299)	(0.883)	(4.429)	(0.837)
Reg Fee * Gender	2.066 **	6.026 **	-0.680	0.183	1.715	-3.498	-4.184 *	10.268	-1.977
	(0.831)	(2.764)	(3.819)	(2.521)	(2.956)	(11.152)	(2.481)	(10.493)	(2.016)
Rsn Maj J Mkt * Gender	0.681	0.430	8.385 ***	3.086 **	28.348 **	7.490 ***	0.775	26.006 ***	3.995 ***
	(0.575)	(0.966)	(2.289)	(1.208)	(12.317)	(1.879)	(2.100)	(6.253)	(1.266)
Rsn Maj S Cont * Gender	-3.900 ***	-1.509	-15.469 ***	-4.974 ***	-9.593 ***	-13.600 ***	-3.384 ***	-48.815 ***	-8.438 ***
	(0.692)	(1.774)	(2.386)	(1.322)	(3.303)	(1.831)	(0.986)	(6.639)	(1.171)
Rsn Maj P Real * Gender	-1.623 ***	1.323 *	-2.170	-0.305	0.416	-1.848	0.174	-7.429 **	-0.239
	(0.576)	(0.786)	(1.398)	(0.815)	(1.189)	(1.168)	(0.785)	(3.790)	(0.732)
Rsn Maj Other * Gender	-0.667	1.412	-1.782	-2.405 *	3.260	-2.097	-0.143	4.530	0.347
	(0.571)	(1.140)	(2.332)	(1.243)	(2.642)	(1.844)	(1.102)	(6.017)	(1.078)
Rsn Univ Free * Gender	-1.156 **	-1.344	-2.565	1.454	-0.219	1.000	-0.893	4.200	-0.799
	(0.520)	(0.880)	(1.719)	(0.994)	(1.613)	(1.372)	(0.916)	(4.517)	(0.835)
Rsn Univ Rep * Gender	-1.163 **	-1.991 **	-1.812	2.309 ***	-0.424	1.622	-0.335	0.463	-0.196
	(0.525)	(0.783)	(1.468)	(0.871)	(1.801)	(1.200)	(0.813)	(3.934)	(0.784)
Rsn Univ Other * Gender	-0.455	-0.201	-1.211	2.763 ***	0.640	1.027	0.545	2.705	-0.279
	(0.496)	(0.823)	(1.657)	(0.915)	(1.686)	(1.341)	(0.845)	(4.351)	(0.854)
Constant	-4.144 ***	-6.121 ***	3.351	3.043 **	1.086	3.502 *	-0.999	83.701 ***	2.383 **
	(0.764)	(1.216)	(2.805)	(1.277)	(2.021)	(1.881)	(1.045)	(10.621)	(1.051)
Sigma	0.213 ***								
	(0.036)								
Number of observations	39,494								

Estimations include the explanatory variables listed in Table 4. * significant at 10%, ** significant at 5%, *** significant at 1%.

TABLE 15. Marginal effects on choice probabilities (Model II)

	Technologies	Exact Sciences	Engineering and Arch.	Natural and Earth Sc.	Arts	Social Sciences	Humanities	Health and Biol. Sc.	Other Health and Biol. Sc.
Gender	0.79 ***	4.52 ***	24.14 ***	-2.81 ***	-1.00 ***	0.07	-3.41 ***	-16.95 ***	-5.36 ***
	(0.15)	(0.22)	(0.41)	(0.27)	(0.18)	(0.32)	(0.19)	(0.41)	(0.25)
White	-0.07	-0.02	-2.11 ***	0.62 **	0.43 **	0.14	-0.05	0.23	0.84 ***
	(0.17)	(0.25)	(0.47)	(0.28)	(0.21)	(0.36)	(0.22)	(0.49)	(0.29)
Work	1.04 ***	1.67 ***	1.25 *	0.59	2.88 ***	2.16 ***	2.12 ***	-12.05 ***	0.33
	(0.24)	(0.37)	(0.70)	(0.43)	(0.39)	(0.61)	(0.33)	(0.68)	(0.42)
Prim Sch Priv	-0.41 *	0.24	0.23	-0.28	0.27	0.42	-0.39	2.04 ***	-2.11 ***
	(0.24)	(0.34)	(0.68)	(0.40)	(0.30)	(0.52)	(0.43)	(0.70)	(0.44)
Prim Sch Mixed	-0.49 **	0.48	-0.48	-0.13	0.72 **	0.45	-0.29	0.90	-1.17 **
	(0.24)	(0.37)	(0.73)	(0.43)	(0.32)	(0.57)	(0.39)	(0.76)	(0.47)
Sec Sch Priv	-0.74 ***	-1.38 ***	-1.59 **	-1.43 ***	-1.62 ***	-1.45 **	0.09	7.83 ***	0.29
	(0.25)	(0.38)	(0.69)	(0.43)	(0.38)	(0.57)	(0.42)	(0.71)	(0.42)
Sec Sch Mixed	-0.45	-0.44	-2.14 *	-1.33 **	-0.10	-1.33	0.71	4.53 ***	0.55
	(0.35)	(0.57)	(1.12)	(0.64)	(0.56)	(0.88)	(0.55)	(1.16)	(0.65)
Sec Sch Tech	0.51 *	1.54 ***	7.78 ***	2.05 ***	-1.09 ***	-1.64 ***	-0.18	-7.77 ***	-1.20 ***
	(0.27)	(0.44)	(0.85)	(0.54)	(0.29)	(0.59)	(0.32)	(0.84)	(0.45)
Other Major	-0.51 **	2.00 ***	2.90 ***	-0.76	0.61	4.26 ***	0.74 **	-7.81 ***	-1.44 ***
	(0.22)	(0.52)	(0.94)	(0.49)	(0.38)	(0.78)	(0.36)	(0.83)	(0.46)
Reg Fee	0.04	5.39 ***	-10.98 ***	-1.69 **	-1.23 ***	0.86	18.67 ***	-15.54 ***	4.49 ***
	(0.26)	(1.33)	(1.10)	(0.67)	(0.31)	(1.77)	(3.30)	(1.51)	(1.34)
Prof Father Non-manual	0.38 *	0.99 ***	0.41	0.66 *	0.37	-0.80 *	0.11	-1.66 ***	-0.48
	(0.22)	(0.32)	(0.56)	(0.36)	(0.25)	(0.43)	(0.35)	(0.57)	(0.36)
Prof Father Manual	0.36	0.92 *	1.57	1.90 ***	0.81 *	-0.94	-0.30	-3.95 ***	-0.36
	(0.29)	(0.48)	(0.99)	(0.61)	(0.45)	(0.74)	(0.43)	(0.98)	(0.54)
Prof Father Other	0.52 *	1.13 ***	0.08	2.12 ***	-0.07	-0.83	0.34	-2.97 ***	-0.31
	(0.29)	(0.43)	(0.83)	(0.55)	(0.33)	(0.60)	(0.42)	(0.82)	(0.49)

TABLE 15. Marginal effects on choice probabilities (Model II, cont.)

	Technologies	Exact Sciences	Engineering and Arch.	Natural and Earth Sc.	Arts	Social Sciences	Humanities	Health and Biol. Sc.	Other Health and Biol. Sc.
Prof Mother Non-manual	-0.09	0.21	-0.17	0.65 *	0.12	-0.21	-0.15	-1.34 **	0.98 ***
	(0.22)	(0.31)	(0.54)	(0.35)	(0.25)	(0.42)	(0.41)	(0.56)	(0.34)
Prof Mother Manual	-0.02	0.02	2.62 *	1.59 *	0.21	-0.87	-0.24	-3.17 **	-0.14
	(0.38)	(0.61)	(1.34)	(0.85)	(0.59)	(1.07)	(0.60)	(1.39)	(0.69)
Prof Mother Housewife	0.28	0.71 **	2.64 ***	0.96 **	-0.38	-1.17 ***	-0.76 *	-3.07 ***	0.79 **
	(0.24)	(0.35)	(0.59)	(0.40)	(0.27)	(0.44)	(0.40)	(0.59)	(0.37)
Prof Mother Other	0.07	0.27	2.02 ***	0.82 *	-0.38	-0.58	-0.05	-3.14 ***	0.97 *
	(0.27)	(0.41)	(0.78)	(0.48)	(0.33)	(0.61)	(0.48)	(0.82)	(0.50)
Rsn Maj Job Mkt	2.78 ***	1.57 ***	4.11 ***	6.12 ***	-3.54 ***	0.10	-2.13 ***	-7.39 ***	-1.62 ***
	(0.38)	(0.45)	(0.79)	(0.63)	(0.16)	(0.59)	(0.26)	(0.77)	(0.42)
Rsn Maj Soc Cont	0.22	-3.16 ***	-18.24 ***	-0.82 **	-2.46 ***	1.80 ***	0.95 ***	21.26 ***	0.44
	(0.23)	(0.28)	(0.54)	(0.40)	(0.22)	(0.52)	(0.33)	(0.69)	(0.39)
Rsn Maj Pers Real	-0.32 **	-0.33	-4.33 ***	-0.01	0.24	-0.40	-0.51 **	4.35 ***	1.31 ***
	(0.15)	(0.27)	(0.49)	(0.30)	(0.24)	(0.35)	(0.21)	(0.49)	(0.31)
Rsn Maj Other	3.11 ***	0.41	-5.25 ***	3.23 ***	-1.73 ***	0.01	0.46	-2.38 ***	2.16 ***
	(0.41)	(0.45)	(0.79)	(0.60)	(0.30)	(0.62)	(0.39)	(0.83)	(0.55)
Rsn Univ Free	0.37 *	-1.05 ***	-2.53 ***	-2.72 ***	-1.79 ***	1.47 ***	0.51 **	4.11 ***	1.64 ***
	(0.20)	(0.29)	(0.55)	(0.33)	(0.25)	(0.43)	(0.26)	(0.58)	(0.36)
Rsn Univ Rep	0.12	-1.03 ***	1.07 **	-2.11 ***	-2.61 ***	1.22 ***	0.85 ***	2.76 ***	-0.27
	(0.19)	(0.26)	(0.48)	(0.31)	(0.21)	(0.36)	(0.23)	(0.49)	(0.30)
Rsn Univ Other	0.76 ***	0.62 *	0.16	-0.47	-2.32 ***	0.52	0.59 **	0.37	-0.24
	(0.22)	(0.32)	(0.56)	(0.38)	(0.23)	(0.42)	(0.24)	(0.56)	(0.33)
Number of observations	39,494								

* significant at 10%, ** significant at 5%, *** significant at 1%.

TABLE 16. Simulated Choice Probabilities (Model II)

	Technologies	Exact Sciences	Engineering and Arch.	Natural and Earth Sc.	Arts	Social Sciences	Humanities	Health and Biol. Sc.	Other Health and Biol. Sc.
Choice probabilities with own entrance probabilities									
Male	2.44 ***	7.41 ***	38.90 ***	7.09 ***	2.74 ***	12.71 ***	2.08 ***	22.41 ***	4.21 ***
Female	1.62 ***	3.01 ***	12.72 ***	9.60 ***	3.73 ***	12.68 ***	6.26 ***	39.91 ***	10.47 ***
Difference	0.83 ***	4.40 ***	26.18 ***	-2.50 ***	-0.99 ***	0.03 ***	-4.18 ***	-17.50 ***	-6.26 ***
Simulated choice probabilities with male entrance probabilities									
Male	2.44 ***	7.41 ***	38.90 ***	7.09 ***	2.74 ***	12.71 ***	2.08 ***	22.41 ***	4.21 ***
Female	1.93 ***	2.89 ***	21.11 ***	12.38 ***	4.00 ***	14.33 ***	7.18 ***	25.02 ***	11.16 ***
Difference	0.51 ***	4.52 ***	17.79 ***	-5.28 ***	-1.26 ***	-1.62 ***	-5.10 ***	-2.61 ***	-6.95 ***
Simulated choice probabilities with female entrance probabilities									
Male	2.43 ***	8.28 ***	28.96 ***	7.57 ***	3.02 ***	13.11 ***	2.43 ***	30.07 ***	4.13 ***
Female	1.62 ***	3.01 ***	12.72 ***	9.60 ***	3.73 ***	12.68 ***	6.26 ***	39.91 ***	10.47 ***
Difference	0.81 ***	5.27 ***	16.24 ***	-2.02 ***	-0.71 ***	0.42 ***	-3.83 ***	-9.84 ***	-6.34 ***

* significant at 10%, ** significant at 5%, *** significant at 1%.

TABLE 17. Marginal effect of gender for interacted variables (Model II)

	Technologies	Exact Sciences	Engineering and Arch.	Natural and Earth Sc.	Arts	Social Sciences	Humanities	Health and Biol. Sc.	Other Health and Biol. Sc.
Average effect	0.79 ***	4.52 ***	24.14 ***	-2.81 ***	-1.00 ***	0.07	-3.41 ***	-16.95 ***	-5.36 ***
	(0.15)	(0.22)	(0.41)	(0.27)	(0.18)	(0.32)	(0.19)	(0.41)	(0.25)
Work	0.95 *	7.20 ***	20.30 ***	-2.69 ***	-1.08 *	-1.73 **	-7.03 ***	-8.10 ***	-7.83 ***
	(0.52)	(0.66)	(0.92)	(0.75)	(0.58)	(0.74)	(0.70)	(0.82)	(0.70)
No Work	0.77 ***	4.02 ***	24.86 ***	-2.83 ***	-0.98 ***	0.40	-2.73 ***	-18.61 ***	-4.89 ***
	(0.14)	(0.22)	(0.45)	(0.28)	(0.18)	(0.35)	(0.18)	(0.46)	(0.27)
Sec Sch Priv	0.64 ***	3.34 ***	25.56 ***	-2.60 ***	-1.02 ***	0.52	-1.96 ***	-19.82 ***	-4.65 ***
	(0.14)	(0.23)	(0.50)	(0.31)	(0.20)	(0.39)	(0.18)	(0.53)	(0.28)
Sec Sch Mixed	0.83	4.12 ***	21.52 ***	-2.06 *	-2.05 **	0.34	-3.89 ***	-12.42 ***	-6.38 ***
	(0.72)	(1.08)	(1.72)	(1.16)	(0.98)	(1.34)	(0.95)	(1.76)	(1.22)
Sec Sch Pub	1.13 ***	7.27 ***	21.34 ***	-3.41 ***	-0.77 **	-1.00 *	-6.64 ***	-11.13 ***	-6.80 ***
	(0.37)	(0.51)	(0.73)	(0.57)	(0.37)	(0.57)	(0.49)	(0.63)	(0.54)
Sec Sch Tech	-0.29	8.06 ***	29.35 ***	-8.85 ***	-1.44 **	-1.98 **	-4.51 ***	-13.45 ***	-6.90 ***
	(0.68)	(0.90)	(1.43)	(1.20)	(0.58)	(0.99)	(0.76)	(1.18)	(0.88)
Sec Sch Not Tech	0.89 ***	4.21 ***	23.68 ***	-2.27 ***	-0.96 ***	0.25	-3.32 ***	-17.26 ***	-5.22 ***
	(0.15)	(0.22)	(0.42)	(0.27)	(0.18)	(0.33)	(0.20)	(0.43)	(0.26)
Other Major	1.55 ***	6.44 ***	17.45 ***	-3.35 ***	-1.83 **	0.38	-3.88 ***	-11.78 ***	-5.00 ***
	(0.55)	(0.86)	(1.20)	(0.86)	(0.75)	(1.06)	(0.88)	(1.19)	(0.80)
No Other Major	0.73 ***	4.35 ***	24.74 ***	-2.76 ***	-0.93 ***	0.04	-3.37 ***	-17.41 ***	-5.39 ***
	(0.15)	(0.23)	(0.43)	(0.28)	(0.18)	(0.33)	(0.19)	(0.43)	(0.27)

TABLE 17. Marginal effect of gender for interacted variables (Model II, cont.)

	Technologies	Exact Sciences	Engineering and Arch.	Natural and Earth Sc.	Arts	Social Sciences	Humanities	Health and Biol. Sc.	Other Health and Biol. Sc.
Reg Fee	2.50 ***	9.14 ***	15.32 ***	0.89	-0.04	-0.39	-10.89 ***	-7.48 ***	-9.04 ***
	(0.68)	(1.03)	(1.22)	(0.91)	(0.58)	(0.95)	(1.15)	(0.90)	(1.01)
No Reg Fee	0.61 ***	4.02 ***	25.09 ***	-3.21 ***	-1.10 ***	0.12	-2.61 ***	-17.97 ***	-4.96 ***
	(0.14)	(0.22)	(0.43)	(0.28)	(0.19)	(0.34)	(0.17)	(0.44)	(0.26)
Rsn Maj Job Mkt	0.62	1.47 *	28.91 ***	-8.38 ***	0.51 **	-2.61 **	-2.88 ***	-13.21 ***	-4.42 ***
	(0.79)	(0.89)	(1.45)	(1.20)	(0.22)	(1.03)	(0.48)	(1.25)	(0.76)
Rsn Maj Soc Cont	0.59	3.65 ***	8.61 ***	2.85 ***	-0.10	5.87 ***	-2.71 ***	-13.28 ***	-5.48 ***
	(0.44)	(0.48)	(0.87)	(0.71)	(0.34)	(0.94)	(0.65)	(1.22)	(0.72)
Rsn Maj Pers Real	0.25	5.83 ***	23.53 ***	-1.41 ***	-0.72 *	0.49	-3.06 ***	-19.46 ***	-5.46 ***
	(0.24)	(0.44)	(0.83)	(0.51)	(0.39)	(0.60)	(0.35)	(0.81)	(0.52)
Rsn Maj Other	1.53 *	6.54 ***	20.13 ***	-5.44 ***	0.09	-1.63	-5.02 ***	-9.42 ***	-6.79 ***
	(0.83)	(0.82)	(1.37)	(1.05)	(0.53)	(1.07)	(0.78)	(1.35)	(1.01)
Rsn Maj Ability	1.01 ***	4.30 ***	27.67 ***	-3.41 ***	-1.74 ***	-0.73	-3.58 ***	-18.31 ***	-5.22 ***
	(0.17)	(0.30)	(0.59)	(0.36)	(0.27)	(0.44)	(0.25)	(0.58)	(0.34)
Rsn Univ Free	0.56	3.95 ***	21.02 ***	-1.44 **	-0.42	1.42 *	-4.19 ***	-14.10 ***	-6.80 ***
	(0.39)	(0.49)	(0.92)	(0.56)	(0.37)	(0.73)	(0.47)	(0.95)	(0.64)
Rsn Univ Rep	0.33	2.51 ***	23.69 ***	-1.26 ***	-0.54 **	1.55 **	-3.10 ***	-19.00 ***	-4.17 ***
	(0.25)	(0.37)	(0.84)	(0.47)	(0.26)	(0.64)	(0.35)	(0.85)	(0.46)
Rsn Univ Other	0.79 **	5.92 ***	22.10 ***	-0.84	-0.62 *	-0.69	-3.53 ***	-17.27 ***	-5.86 ***
	(0.39)	(0.54)	(0.92)	(0.65)	(0.35)	(0.71)	(0.46)	(0.89)	(0.57)
Rsn Univ Best for Course	1.16 ***	5.32 ***	26.58 ***	-5.11 ***	-1.67 ***	-1.03 **	-3.21 ***	-16.84 ***	-5.22 ***
	(0.21)	(0.34)	(0.62)	(0.44)	(0.33)	(0.46)	(0.27)	(0.61)	(0.38)
Number of observations	39,494								

* significant at 10%, ** significant at 5%, *** significant at 1%.